DOCUMENT RESUME

ED 238 940 TM 840 042

AUTHOR Wilcox, Rand R.
TITLE Optimal Measurement Considerations for Diagnostic Tests. Methodology Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 83
GRANT NIE-G-83-0001
NOTE 108p.
PUB TYPE Collected Works - General (020) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Diagnostic Tests; *Estimation (Mathematics); Guessing (Tests); *Latent Trait Theory; *Measurement Techniques; *Multivariate Analysis; Scoring; Testing Problems; Test Items; *True Scores
IDENTIFIERS Linear Measurement

ABSTRACT

This document presents a series of five papers describing issues in educational measurement. "A Simple Model for Diagnostic Testing When There Are Several Types of Misinformation" directly addresses the diagnostic issue. It describes a simple latent trait model for testing, examines use of erroneous algorithms, and illustrates the derivation of an optimal scoring rule for multiple choice test items. "Measuring Mental Abilities with Latent State Models" has three goals: to review the latent state models that have been proposed for measuring aptitude and achievement; to outline the measurement problems that can now be solved with latent state models; and to discuss how latent state and latent trait models are related. "Strong True Score Theory" reviews true score models in light of various assumptions about guessing. "Approximating Multivariate Distributions" suggests a simple approximation of multivariate distributions. The suggested method is compared with several other approximations. These comparisons indicate that the new approximation nearly always gives better results. "Unbiased Estimation in a Closed Sequential Testing Procedure" provides an optimal linear estimator of the proportion of items within an item domain that an examinee would answer correctly if every item were attempted. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED238940

Deliverable - November 1983

METHODOLOGY PROJECT

OPTIMAL MEASUREMENT CONSIDERATIONS FOR DIAGNOSTIC TESTS

Rand R. Wilcox

Grant Number

NIE-G-83-0001

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Gray

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

November 1983

TM 840 042

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

PREFACE

This document presents a series of papers describing issues in educational measurement. The first, paper, "A Simple Model for Diagnostic Testing When There Are Several Types of Misinformation," directly addresses the diagnostic issue. It describes a simple latent trait model for testing, examines use of erroneous algorithms, and illustrates the derivation of an optimal scoring rule for multiple choice test items.

The second paper, "Measuring Mental Abilities with Latent State Models," has three goals: 1) to review the latent state models that have been proposed for measuring aptitude and achievement; 2) to outline the measurement problems that can now be solved with latent state models, and 3) to discuss how latent state and latent trait models are related.

The third paper, "Strong True Score Theory," reviews true score models in light of various assumptions about guessing. It is an invited paper to appear in an encyclopedia for statistics.

The fourth paper, "Approximating Multivariate Distributions," suggests a simple approximation of multivariate distributions. The suggested method is compared with several other approximations. These comparisons indicate that the new approximation nearly always gives better results.

The final paper, "Unbiased Estimation in a Closed Sequential Testing Procedure" provides an optimal linear estimator of the proportion of items within an item domain that an examinee would anwer correctly if every item were attempted.

A SIMPLE MODEL FOR DIAGNOSTIC TESTING WHEN THERE ARE SEVERAL TYPES OF MISINFORMATION

Rand R. Wilcox
Department of Psychology
University of Southern California

and

Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

In diagnostic testing one purpose of a test might be to determine whether an examinee has acquired the appropriate skills for solving certain types of problems, or whether the examinee is using an erroneous algorithm. In the latter case it is also desired to determine which of several erroneous algorithms is being used so that remedial training can be given. Birenbaum and Tatsuoka (1982) recently illustrated that when testing eighth graders on the addition of signed numbers, examinees might indeed be applying one of several erroneous algorithms, and more recently they reported results on a scoring procedure for this situation. This paper describes a simple latent class model for handling the items in Birenbaum and Tatsuoka; included is a description and illustration of how to derive the optimal scoring rule when multiple choice test items are used.

Birenbaum and Tatsuoka (1982) provide an interesting example of the need to measure and classify examinees according to the type of misinformation they have relative to a particular skill. They were specifically concerned with testing the addition of signed numbers, but it is evident that similar problems occur in many situations. As Birenbaum and Tatsuoka point out, examinees might be using one of several erroneous algorithms when responding to these items. They described three algorithms that were actually used by examinees, and since they play an important role here, they are briefly reviewed.

The first erroneous algorithm was treating parentheses as meaning absolute value. Thus 7+(-3) would result in an answer of 10. The second algorithm was to add the two numbers and take the sign of the number having the largest absolute value. For example, if asked to compute 3+ -7, the examinee adds 3 and 7, and because 7>3, a negative sign is added yielding -10. The third erroneous algorithm was to add the two numbers when they had different signs, and to put a plus sign in the result. For example, 3+ -7=10 according to this rule. If the two numbers have the same sign, the student takes their difference and puts the common sign in the result. For example, (-8)+(-4) =-4. This last algorithm resulted from the student misunderstanding how to use the number line as it was explained by the teacher. Table 1, taken from Birenbaum and Tatsuoka (1982) shows several addition problems and the results arrived at according to the three erroneous algorithms just described. Note that different algorithms can yield the same answer, and in some cases even the correct response.

Birenbaum and Tatsuoka (1982, 1983) argue for the need to measure misinformation and to determine the type of misinformation that a student has. In their more recent article (Birenbaum & Tatsuoka, 1983), they compared two scoring algorithms for measuring misinformation, but no results were given on determining the accuracy of either procedure, and indeed neither procedure was developed with the goal of finding the optimal scoring procedure for identifying whether an erroneous algorithm is being used. (They compared coefficient alpha for the two scoring procedures, but this is not a direct measure of the accuracy of the test as it is defined below.)

The goal in this paper is to illustrate how an optimal scoring procedure can be derived for the situation considered in Birenbaum and Tatsuoka (1982). As will become evident, the process used for determining the optimal scoring rule can be easily extended to other situations, but to keep the illustration as simple as possible, attention will be restricted to the items in Table 1. An additional advantage of the results to be given is that expressions are also derived for the probability of correctly determining the algorithm being used by an examinee.

Before continuing, some comments should be made regarding results similar to the developments made here. First, the problem being examined is similar to one considered by Macready and Dayton (1977). It is easily seen though that the latent structure model they used is inappropriate for the problem at hand. Wilcox (1982a) proposed a model for measuring misinformation via an answer-until-correct scoring procedure, but this model is inadequate here as well. The reason is that his model can measure only one

type of misinformation, and here the problem is contending with three erroneous algorithms. Dayton and Macready (1980) as well as Goodman (1974) describe very general latent class models that could be applied, and Bergan et al. (1980) described an appropriate scoring procedure. However, these models require iterative techniques that may be unnecessarily complicated. In particular, Dayton and Macready's model requires iterative approximations of the maximum likelihood estimates of the parameters, and for theoretical reasons it is best to avoid these estimation techniques whenever possible (Kale, 1962a; 1962b). The problem is determining whether iterative estimation procedures converge to the maximum likelihood estimates that they are intended to approximate. It appears that they usually do, but there is no guarantee that this will always be the case. (For a situation where iterative techniques can converge to inappropriate values, see Wilcox, 1979.) Thus, an important aspect of this paper is that by making certain assumptions about how examinees behave when taking test items, which are motivated by a published empirical study described below, a relatively simple model results where explicit maximum likelihood estimates of the parameters are available, and these estimates can be used to solve the measurement problems described above.

2. The Model and Its Assumptions

It is assumed that multiple-choice test items are used, and that every item has t alternatives. This last assumption is made primarily for notational convenience. Using multiple choice items introduces the problem of guessing, but this seems to be easier to handle, from a statistical point of view,

than is the problem of careless errors which is one of the erroneous algorithms also considered by Birenbaum and Tatsuoka (1982). Here it is assumed that careless errors occur with probability close to zero so that for practical purposes this error can be ignored. As explained in the introduction, only the three erroneous algorithms in Birenbaum and Tatsuoka will be considered, plus, of course, the algorithm of random guessing. Thus, for the population of examinees to be tested, it is assumed that every examinee belongs to one of five mutually exclusive latent states: they know how to solve the items, they guess at random, or they apply one of the three incorrect algorithms described above. It is also assumed that if an examinee is using the correct algorithm, the correct response is always chosen, and if one of the three erroneous algorithms is used, an examinee will always choose a corresponding response. For comments about this last assumption, see section 6. For the moment it is also assumed that every item has a distractor that is consistent with each of the erroneous algorithms. This restriction could be relaxed, if desired, when applying the procedure outlined in section 5. Another assumption is that there are no examinees who have partial information. Although empirical results indicate that partial information exists in some situations (e.g., Coombs et al. 1956), there is also some empirical evidence that when dealing with misinformation, it may be reasonable to assume that no examinees have partial information (Wilcox, 1982a). It is not being suggested that this assumption be taken for granted, only that it might be reasonable in practice--section 4 dis-

cusses how certain implications of the model can be tested, and this test should always be carried out.

The next step is to find n items that make it possible to distinguish between any two examinees having a different erroneous algorithm. In addition, these items should include at least one item that will result in at least one incorrect response when an erroneous algorithm is being used. These last two conditions are clearly satisfied for the items in Table 1. In fact only the first three items are needed.

Let a 1 and 0 represent a correct and incorrect response to an item, respectively. Consider an examinee responding to the first three items in Table 1. If the first erroneous algorithm is used, the resulting response pattern will be (1,0,1). For the second erroneous algorithm, the response pattern will be (0,0,1), and the third erroneous algorithm will give (0,0,0). Thus, if an examinee has response pattern (1,0,1), for example, the assumptions of the model rule out the possibility that the examinee is using one of the other two erroneous algorithms, and so the examinee is either applying the first erroneous algorithm or is guessing at random.

Two practical problems will be considered. The first is estimating the proportions of examinees among a population of examinees who belong to the various latent states. The second problem is distinguishing between those examinees who are guessing at random, and those belonging to one of the four other latent classes. As will become evident, a solution to the first problem can be useful when solving the second.

Let ζ be the proportion of examinees who know the correct algorithm, and let ζ_i (i=1,2,3) be the proportion who are using the ith erroneous algorithm. Finally, let ζ_4 be the proportion of examinees who guess at random, and let p_{ijk} (i=1,1; j=0,1; k=0,1) be the probability that a randomly sampled examinee would give the response pattern (i,j,k). For example, p_{101} is the probability that a randomly sampled examinee would give a correct, incorrect and correct response to the first three items in Table 1. From the assumptions already described, it follows that

$$p_{111} = \zeta + \zeta_4(1/t)^3 \qquad (2.1)$$

$$p_{101} = \zeta_4(1/t)^2(1-1/t)+\zeta_1 \qquad (2.2)$$

$$p_{001} = \zeta_4(1/t)(1-1/t)^2+\zeta_2 \qquad (2.3)$$

$$p_{000} = \zeta_4(1-1/t)^3+\zeta_3 \qquad (2.4)$$

$$p_{110}=p_{011}= \zeta_4(1/t)^2(1-1/t) \qquad (2.5a)$$

$$p_{100}=p_{010} = \zeta_4(1/t)(1-1/t)^2 \qquad (2.5b)$$

For N randomly sampled examinees, let x_{ijk} be the number of examinees having response pattern (i,j,k), and let q_1 be the common value of $p_{110}=p_{011}$ and let $q_2=p_{100}=p_{010}$ be the common value of $p_{100}=p_{010}$. From standard results on the multinomial distribution in conjunction with results in Zehna (1966), $\hat{q}_1=(x_{110}+x_{011})/2N$ and $\hat{q}_2=(x_{100}+x_{010})/2N$ are maximum likelihood estimates of q_1 and q_2. It follows that maximum likelihood estimates of $\zeta_4,\zeta_3,\zeta_2,\zeta_1$ and ζ are

$$\hat{\zeta}_4=(\hat{q}_1(1/t)^{-2}(1-1/t)^{-1}+\hat{q}_2(1/t)^{-1}(1-1/t)^{-1})/2 \qquad (2.6)$$

$$\hat{\zeta} = x_{111}/N - \hat{\zeta}_4/t^3 \tag{2.7}$$

$$\hat{\zeta}_1 = x_{101}/N - \hat{\zeta}_4(1-1/t)/t^2 \tag{2.8}$$

$$\hat{\zeta}_2 = x_{001}/N - \hat{\zeta}_4(1-1/t)^2/t \tag{2.9}$$

and

$$\hat{\zeta}_3 = x_{000}/N - \hat{\zeta}_4(1-1/t)^3 \tag{2.10}$$

Making Decisions About an Examinee's Latent State

Suppose an examinee gives the response pattern (1,1,1). Then according to the model, the examinee is either using the correct algorithm or choosing responses at random; the problem is determining which is true. The simplest solution is to examine the probability of observing the response pattern (1,1,1) if the examinee is guessing at random; this is just t^{-3}, assuming the responses are independent of one another. If t^{-3} is small, it might be decided that the examinee is not guessing, but this approach can be unsatisfactory. To illustrate why, suppose $\zeta=0$. Then the optimal scoring rule would be to always decide an examinee does not know, and to conclude therefore that an examinee is guessing at random when the response pattern (1,1,1) is given.

The question arises as to whether the optimal rule for $\zeta=0$ is also optimal when $\zeta>0$, and if so, how far away from zero can ζ be before some other rule should be used. There is also the problem of determining the overall accuracy of the decision rule being used. A solution to the first problem is to decide an examinee is using the correct algorithm if and

only if the response pattern (1,1,1) is given and

$$\zeta_4 t^{-3} < \zeta. \tag{2.11}$$

This rule is derived by noting that the joint probability of randomly sampling an examinee who guesses at random and who gives the response (1,1,1) is just $\zeta_4 t^{-3}$. Also, the joint probability of sampling an examinee who knows and who gives the response pattern (1,1,1) is ζ. Thus, if $\zeta < \zeta_4 t^{-3}$, decide the examinee is guessing at random. Optimal properties of this decision rule (given in a more general context) are described by Copas (1974).

A similar approach can be used to derive decision rules for determining whether an examinee is using a particular erroneous algorithm. Suppose, for example, the response (1,0,1) is given. Then according to the model, the examinee either is using the first erroneous algorithm, or is guessing at random. The optimal rule is to decide the examinee is using the erroneous algorithm if and only if

$$\zeta_1 > \zeta_4 [t^{-2}(1-t^{-1})] \tag{2.12}$$

where $t^{-2}(1-1/t)$ is the probability of the response pattern (1,0,1) from an examinee guessing at random. Thus, this is the same rule as (2.11) except that ζ has been replaced by ζ_1 and t^{-3} has been replaced by the probability of the response pattern (1,0,1) from an examinee guessing at random. Similar modifications are made for the other two response patterns corresponding to the other two erroneous algorithms. As for the response patterns corresponding to equations (2.5), simply decide the examinee is guessing at random.

Extensions to n Item Tests Having k Latent States

The basic process used to analyze the first three items in Table 1 is easily extended to n item tests involving k latent states. Consider any response pattern A where A is a vector of 1's and 0's. Let C_k be the joint probability of observing A and having an examinee in the kth latent state, k=1,...,K. Then decide that an examinee is in the ith latent state if $C_i=\max C_k$. Another illustration is given in section 5. As already mentioned, when trying to classify an examinee as belonging to one of two latent states, this rule is known to have certain optimal properties (Copas, 1974). If an examinee giving response A can belong to more than two latent classes, the rule used here is the same as the one used by Bergan et al. (1980), but the optimal properties discussed by Copas (1974) have not been established.

3. The Probability of a Correct Decision

Returning to the analysis of the first three items in Table 1, suppose the procedure in the previous section has been applied, and that a scoring rule has been determined. The next problem is determining whether there is a high likelihood of making a correct decision about the latent state of a randomly sampled examinee. If this probability is judged to be too low, the test might be modified as described below. Again to explicate the process, only the K=5 latent states of section 2 will be considered.

Suppose the decision rule in Table 2 is to be used. Then for a randomly sampled examinee, the probability of a correct decision (PCD) is just

$$\zeta + \zeta_1 + \zeta_2 + \zeta_3 + B\zeta_4 \qquad (3.1)$$

where

$$B = 2t^{-2}(1-t^{-1})+2t^{-1}(1-t^{-1})^2 .$$

Suppose instead that for response pattern (0,0,0) it is decided an examinee is guessing at random. Then (3.1) becomes

$$\zeta + \zeta_1 + \zeta_2 + C\zeta_4 \qquad (3.2)$$

where $C=B+t^{-3}$. Similar adjustments can be made if the decision rule in Table 2 is modified in any way.

The general technique in determining an expression for the PCD is to first derive an expression for the probability that, for an examinee guessing at random, the observed response pattern will correspond to one where the decision is made that the examinee is indeed guessing at random. Consider, for example, response pattern (1,1,0). Given that the examinee is guessing at random, the probability of this response pattern is $(t^{-1})(t^{-1})(1-t^{-1})$. Repeating this process for every response pattern for which it is decided that an examinee is guessing at random and adding the results yields the coefficient for ζ_4 in the expression for the PCD. For the decision rule in Table 2, there are four such response patterns, and they add to B in (3.1). For examinees in the other latent states, the response pattern is determined with probability one, and so no coefficients

are needed for them in (3.1). If the PCD is judged to be too small, additional items or more distractors can be used, and then the process described above is applied again.

4. Comments About Testing the Model

A partial check on the model in section 2 is to test (2.5) with the usual sign test. In the more general case, such as in section 5, it is necessary to test for equal cell probabilities among several cells, and this is usually accomplished with a chi-square test. The purpose of this section is to make some brief comments about this well known procedure.

First, exact tests for equiprobable cells can be made when N, the number of examinees, is less than or equal to 50 (Smith et al., 1979; Katti, 1973). In cases where a chi-square distribution must be used to get approximate critical values, it appears that a better approximation of the critical values can be had by applying results in Wilcox (1982b).

Second, a practical problem with testing for equiprobable cells is that the null hypothesis might be rejected even when the cell probabilities are nearly equal in value. Of course this is particularly likely to happen when N is large. Accordingly, if the chi-squre test is significant, it would seem prudent to estimate the overall inequality among the cell probabilities, and a detailed discussion about how this can be done can be found in Wilcox, Cliff and Embretson (to appear).

5. An Alternative Approach

In some cases it may be useful to take into account the actual response chosen by an examinee rather than limiting the analysis to the pattern of correct and incorrect responses. By doing this, fewer items may be required in order to obtain an accurate test.

As an illustration, suppose items 1 and 5 in Table 1 are to be used. If the observed responses are -4 and -32, respectively, the examinee was either guessing at random or was using the correct algorithm. Thus $Pr(-4,-32)=\zeta+\zeta_4(t^{-1})(t^{-1})$. As a more specific example, suppose there are t=3 alternatives for both items 1 and 5 in Table 1. Then

$$\xi_1 = Pr(-4, -32) = \zeta+\zeta_4/3^2$$

where the symbol ξ is introduced for notational convenience. In a similar manner the probability of all possible response patterns can be written in terms of the ζ's and they are

$$\xi_2=Pr(-4,32)=\zeta_1+\zeta_4/9$$

$$\xi_3 = Pr(-4,-14)=\zeta_4/9$$

$$\xi_4 = Pr(-10,32)=\zeta_4/9$$

$$\xi_5 = Pr(-10,-32)=\zeta_2+\zeta_4/9$$

$$\xi_6= Pr(-10,-14)=\zeta_4/9$$

$$\xi_7= Pr(10,32)=\zeta_4/9$$

$$\xi_8= Pr(10,-32)=\zeta_4/9$$

$$\xi_9= Pr(10,-14)=\zeta_3+\zeta_4/9$$

Observe that $\xi_3 = \xi_4 = \xi_6 = \xi_7 = \xi_8$, and so the comments in section 4 apply.

Let $\hat{\xi}_i$ be the usual maximum likelihood estimate of ξ_i. Then the set of equations just given imply that

$$\hat{\zeta}_4 = 9(\hat{\xi}_3 + \hat{\xi}_4 + \hat{\xi}_6 + \hat{\xi}_7 + \hat{\xi}_8)/5$$

is a maximum likelihood estimate of ζ_4. Hence

$$\hat{\zeta}_3 = \hat{\xi}_9 - \hat{\zeta}_4/9$$

$$\hat{\zeta}_2 = \hat{\xi}_5 - \hat{\zeta}_4/9$$

$$\hat{\zeta}_1 = \hat{\xi}_2 - \hat{\zeta}_4/9$$

and

$$\hat{\zeta} = \hat{\xi}_1 - \hat{\zeta}_4/9$$

are maximum likelihood estimates of ζ_3, ζ_2, ζ_1 and ζ respectively. Thus, only two items were needed to estimate the proportion of examinees in the five latent states.

Next suppose the ζ's are known or that they have been estimated, and that a scoring procedure must be established. Consider in particular the response pattern (-4, -32). The joint probability of using the correct algorithm and giving the response pattern (-4, -32) is just ζ. The joint probability of guessing at random and giving the response pattern (-4, -32) is $\zeta_4(t^{-1})(t^{-1}) = \zeta_4/9$. Thus, for the response pattern (-4, -32), if $\zeta > \zeta_4/9$ decide an examinee is using the correct algorithm. If $\zeta < \zeta_4/9$, decide the examinee is guessing at random. The important point here is that

the analysis is basically the same as it was in section 2. Of course the other response patterns can be analyzed in the same manner. An expression for the PCD can also be determined once a scoring rule has been settled upon. The details are basically the same as before, and so further comments are omitted.

6. Concluding Remarks

In section 2 it was assumed that every item has an alternative that is consistent with at least one of the algorithms that might be used by an examinee. It should be noted that if computerized testing is possible, an adaptive test could be administered that relaxes this assumption. Suppose, for example, an item is given and that the observed response rules out the possibility that an examinee is using the first erroneous algorithm. Then the next item could be chosen based on the assumption that the examinee is not using this algorithm. That is, the distractors need not include an alternative that is consistent with the first erroneous algorithm. When measuring complex skills, this approach could be important.

One of the assumptions of the model was that there is no carelessness. That is, if an examinee is using a particular algorithm to arrive at an answer, the alternative corresponding to this algorithm will always be chosen. In some cases it might be necessary to include the possibility that an examinee might carelessly choose an alternative that is inconsistent with the algorithm being applied. The models used here are easily extended to

handle this problem, but iterative estimates of the parameters would be needed. One way to solve this estimation problem is to proceed as outlined in Goodman (1979). Once the parameters are estimated, a scoring rule can be derived as was outlined above.

Another important point is that the scoring rules described here are based on the assumption that the goal is to maximize the number of examinees for whom a correct decision is made about their latent state. This could mean, however, that an examinee could get an item right, and yet it would still be concluded that an erroneous algorithm was being used. If this possibility is objectionable, some other scoring rule should be considered. However, the results given here are still valuable because they yield a method of assessing the accuracy of a test if a conventional scoring rule is applied, and the scoring rule described here might be useful when evaluating the effectiveness of a particular instructional program.

TABLE 1

Problems and Responses According to the Three Erroneous Algorithms in Birenbaum and Tatsuoka (1982)

Problem No.	Erroneous Algorithm 1	2	3
1. 3 + -7 = -4	-4	-10	10
2. 7 + (-3) = 4	10	10	10
3. -6 + -15 = -21	-21	-21	-9
4. -6 + +15 = 9	9	21	21
5. (-23) + (-9) = -32	32	-32	-14

TABLE 2

A Decision Rule for the First Three Items in Table 1

Response Pattern of Corrects and Incorrects	Decision
111	Uses the correct algorithm
110	Guessing at random
101	Uses the first erroneous algorithm
011	Guessing at random
100	Guessing at random
010	Guessing at random
001	Uses the second erroneous algorithm
000	Uses the third erroneous algorithm

References

Bergan, J. R., Cancelli, A. A., & Luiten, J. W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. Journal of Educational Statistics, 1980, 5, 65-81.

Birenbaum, M., & Tatsuoka, K. On the dimensionality of achievement test data. Journal of Educational Measurement, 1982, 19, 259-266.

Birenbaum, M., & Tatsuoka, K. The effect of a scoring system based on the algorithm underlying the student's response patterns on the dimensionality of achievement test data of the problem solving type. Journal of Educational Measurement, 1983, 20, 17-26.

Coombs, C. H., Dholland, J. E., & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

Copas, J. B. On symmetric compound decision rules for dichotomies. Annals of Statistics, 1974, 2, 199-204.

Dayton, C. M., & Macready, G. B. A scaling model with response errors and intrinsically unscalable respondents. Psychometrika, 1980, 45, 343-356.

Goodman, L. A. On the estimation of parameters in latent structure analysis. Psychometrika, 1979, 44, 123-128.

Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 1974, 61, 215-231.

Kale, B. K. On the solution of likelihood equations by iteration processes. The multiparametric case. Biometrika, 1962, 49, 479-486. (a)

Kale, B. K. A note on a problem in estimation. Biometrika, 1962, 49, 553-557. (b)

Katti, S. K. Exact distribution for the chi-square test in the one way table. Communications in Statistics, 1973, 2, 435-447.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 199-120.

Smith, P. J., Rae, D. S., Manderscheid, R. W., & Silbergeld, S. Exact and approximate distributions of the chi-square statistic for equiprobability. Communications in Statistics -- Simulation and Computation, 1979, B8, 131-149.

Wilcox, R. R. Estimating the parameters of the beta-binomial distribution. Educational and Psychological Measurement, 1979, 31, 527-535.

Wilcox, R. R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74. (a)

Wilcox, R. R. A comment on approximating the X^2 distribution in the equiprobable case. Communication in Statistics -- Simulation and Computation, 1982, 11, 619-623. (b)

Wilcox, R., Cliff, N., & Embretson, S. Measuring mental abilities: Advances in statistical theories. Beverly Hills: Sage Publishing Co., to appear.

Zehna, P. W. Invariance of maximum likelihood estimation. Annals of Mathematical Statistics, 1966, 37, 744.

MEASURING MENTAL ABILITIES WITH LATENT STATE MODELS

Rand R. Wilcox
Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

The three goals in this paper are (1) to review the latent state models that have been proposed for measuring aptitude and achievement, (2) to outline the measurement problems that can now be solved with latent state models, and (3) to discuss how latent state and latent trait models are related. It is pointed out that latent state and latent trait models measure different things that are related to one another in a complicated fashion.

1. INTRODUCTION

There are now four interrelated approaches to measuring aptitude and achievement that are based on different notions of true scores. Classical test theory is the best known approach where ability is defined in terms of a propensity distribution. The other three are latent trait models, item sampling models, and latent state models. No doubt latent state models are the least well known. One reason for this is that early models made very restrictive or inconvenient assumptions, and even if the models could be applied, it was unclear how to solve the many measurement problems that arise in practice (cf. Meskauskas, 1976).

Today the situation has changed radically, there are now latent state models that are relatively easy to use, and empirical investigations indicate that the underlying assumptions are usually met, or that they are reasonable approximations of reality. Just as important is that many measurement problems can be solved that were previously impossible to address. The three major goals in this paper are to (1) review the various latent state models, (2) describe some of the measurement problems that can now be solved with latent state models, and (3) briefly indicate how latent trait models, item sampling models, and latent class models are related to one another. The last goal is particularly important because when there are errors at the item level such as guessing, all three models estimate different quantities that are related to one another in a complicated fashion. In fact, if a measurement problem is formulated in terms of one model it may be very difficult to find a satisfactory reformulation of

the problem in terms of another model. This point is elaborated below. Accordingly, it is important to consider the differences among the models when addressing a particular measurement problem.

It should be stressed that none of the models described below are considered to be always bad or inappropriate. The position advocated here is that an eclectic approach to measuring mental abilities should be used. That is, the choice of a true score model should be dictated, at least in part, by the goal of the test, or the type of ability being estimated. All that is being suggested is that different models are based on different constructs, and so they estimate different things, which suggests that some models may be inappropriate in some situations, or that several models might be used to study a test. For example, the type of guessing examined in latent state models is completely ignored in all other models, and so if this type of guessing is deemed important, a latent state model should be used. There is a widespread belief that the guessing parameter in latent trait models is the same as the notion of guessing in latent state models, but this is not true. In section 6 an attempt is made at explaining the difference.

The paper is organized as follows: Section 2 briefly reviews the basic elements of latent trait models that will be needed in the paper. Section 3 does the same for item sampling models, and some comments are made about how these models relate to latent trait models. Section 4 reviews the theoretical developments in latent class models that are specifically intended for measuring aptitude and achievement. Certain

aspects of these models were reviewed by Macready and Dayton (1980b) and so these features will not be discussed here. Section 5 describes applications that can not be addressed by other measurement models. Included are generalizations of item sampling models. Section 6 makes additional comments on how latent trait and latent class models are related to one another. In particular, this section discusses the importance of guessing in latent trait models, and it points out that the type of guessing examined in latent class models is completely ignored in latent trait models--even in Birnbaum's three parameter model.

2. Latent Trait Models

Latent trait models are discussed in detail by Birnbaum (1968), Lord and Novick (1968, ch. 16), Lord (1980), and Hambleton et al. (1978) give an excellent review of this approach to mental test theory. See also, the 1977 special issue of the Journal of Educational Measurement, Weiss and Davison (1981), and the 1982 special issue of Applied Psychological Measurement.

Generally, these models express the probability of an examinee giving a correct response to an item as a function of an examinee's "ability" and certain item parameters. For example, the Rasch model postulates that $p(\theta)$, the probability a specific examinee with ability level θ $(-\infty < \theta < \infty)$ will produce the correct response to a dichotomously scored item, is

$$p(\theta) = \exp(\theta - b)/(1 + \exp(\theta - b)) \tag{2.1}$$

where b (the difficulty level) is a parameter that characterizes the item. (See, for example, Wright, 1977; Wainer et al., 1980.)

An alternative expression for $p(\theta)$ is the two parameter normal ogive model given by

$$p(\theta) = \int_{-\infty}^{L} \phi(t)dt \tag{2.2}$$

where $\phi(t)$ is the standard normal probability function, $L = a(\theta - b)$, and a is the item "level of discrimination". A closely related model is the two parameter logistic model where

$$p(\theta) = (1 + \exp(-1.7a(\theta - b)))^{-1} \tag{2.3}$$

(Birnbaum, 1968, p. 400). An even more general three parameter model is given by

$$p(\theta) = c + (1 - c)\frac{\exp(1.7a(\theta - b)}{1 + \exp(1.7a(\theta - b))} \tag{2.4}$$

where c is the probability of a correct response from an examinee with low ability. In all of the above models, the symbols a, b, and c represent unknown parameters that are estimated with the observed scores of a sample of examinees. A particularly important feature of latent trait models is that once the item parameters are estimated, it is possible to construct a test so that the expected observed scores will have certain properties that are deemed important.

Numerous articles on latent trait models have been published. However, as previously indicated, the goal of this paper is not to summarize these results. For present purposes, the important point is the interpretation of $p(\theta)$. One interpretation is that p is the probability of a correct response over repeated independent administrations of the item. In other words, p is the examinee's expected observed score, where the expectation is defined in terms of a propensity distribution. However, Lord (1980, ch. 15; 1974) argues that this interpretation leads to certain logical problems, and so he proposes that one of two other interpretations be used instead. The first imagines a pool of items all of which have the same item parameters a, b, and c. Then $p(\theta)$ is the probability that a specific examinee with ability θ will give the correct response to an item randomly sampled from this item domain. The actual items on a test will typically have different item parameters, and so each of these items would be viewed as being sampled from an item domain corresponding to the values of a, b, and c.

The second interpretation views examinees, rather than items, as being randomly sampled. For an item with parameters a, b, and c, $p(\theta)$ is the probability of a correct response from a randomly sampled examinee who has ability level θ.

Some other basic assumptions associated with latent trait models should be mentioned. One of these is the assumption of local independence. This means that given θ, responses are independent of one another. Letting p_i be the value of $p(\theta)$ for the ith item on a test, local independence means that if items are scored dichotomously, the probability of y items correct given θ is

$$f(y|\theta) = \sum \prod_{i=1}^{n} p_i^{x_i} (1 - p_i)^{1-x_i} \qquad (2.5)$$

where $x_i = 1$ or 0 according to whether the ith item is answered correctly, and where the summation is over all vectors $(x_1, \ldots, x_n)$ such that $\Sigma x_i = y$. A test of this assumption was recently proposed by Holland (1981), but it has not yet been applied to real data.

Another property of the most commonly used latent trait models is that they are unidimensional. This means that only one person parameter, namely θ, is needed to determine the probability of a correct response to an item. McDonald (1981) points out that latent trait models can be viewed as a nonlinear factor analysis model with only one factor (cf. Mellenbergh, 1981).

Another observation will be useful later. This is that if all the items on an n-item test have the same item parameters (i.e., the same values for a, b, and c) then (2.5) reduces to

$$f(y|\theta) = \binom{n}{y} p^y (1 - p)^{n-y} \qquad (2.6)$$

the binomial probability function, where p is the common value of the p_i's.

Finally, for multiple choice items, latent trait models do not deal with the construct "knowing" in any way -- they deal with the probability of a correct response which is different from the probability of knowing.

3. Item Sampling Models

A third class of true score models is known as item sampling models. The binomial error model is the one most frequently used; a recent review is given by Wilcox (1981a), and so only its basic properties will be given here.

Consider a single examinee responding to an n-item test. One situation leading to the binomial error model is where the n items are actually sampled from some larger item domain. If ξ is the proportion of items the examinee would get correct if every item in the item pool were attempted, then the probability of y correct responses is

$$f(y|\xi) = \binom{n}{y} \xi^{y}(1 - \xi)^{n-y} \qquad (3.1)$$

(It is assumed that sampling is from an infinite pool, or finite pool with replacement, and that ξ remains constant over the trials.) In many situations items are not randomly sampled, and there is no item pool. Thus, there is no a priori reason for assuming (3.1) holds. It might seem, therefore, that the binomial error model is not really justified, but the point is that (3.1) might give a good fit to data. Indeed, the empirical investigations cited by Wilcox (1981a) suggest that (3.1) will frequently give good results when addressing various measurement problems. Note that there is also no a priori reason for using latent trait models (Lord, 1980). Again the crucial question is whether the models give good results with real data.

It might appear that the binomial error model is more restrictive than latent trait models in the sense that if the item parameters a, b, and c are the same for every item, the probability of y correct responses

is given by (2.6) which is the same form as (3.1). In particular, one might conclude that ξ in (3.1) and p in (2.6) are the same. They are related but in a more complicated fashion.

Typically, the n items on a test will have different values for a, b, and c. If items are really sampled from some item domain, the corresponding item parameters will have some distribution, say g(a,b,c). Thus, for a randomly sampled item, the probability of a correct response from a specific examinee with ability level θ is $\xi = E(p(\theta))$ where the expectation is taken with respect to the random variables a, b, and c. That is, $\xi = \iiint p(\theta)\, g(a,b,c)\, da db dc$.

To illustrate the practical implications of this result, consider a criterion-referenced test where the goal is to determine whether an examinee's percent correct true score ξ is above or below the known constant ξ_0. Is it possible to formulate the problem in terms of a latent trait model? In particular, how can a criterion score be found (a value of θ_0) that corresponds to ξ_0. If the suggestion in Lord (1980, p. 174) is followed, one might determine the criterion score to be the value of θ such that

$$n\xi_0 = \sum_i p_i(\theta) \qquad (3.2)$$

where $p_i(\theta) = p(\theta, a_i, b_i, c_i)$ is the item response function for the ith item on an n-item test. The point is that if a different set of items were used with presumably different item parameters, equation (3.2) would yield a different criterion score. Thus, this procedure yields, at best, an estimate of what the criterion score would be if the problem were to be reformulated in terms of θ_0.

Observe that $n\xi_0$ is different from the true score used by Lord (1980, p. 174). Lord is referring to an expected number-correct true score, but

the expectation is different from $n\xi_0$ in equation (3.1), as was explained above.

Does this mean that one model is better than another? The answer is an unequivocal no; the point is that they are not exactly the same, and the choice of a model should depend on what an investigator wants to know. Of course, some individuals might be dissatisfied with both models. In terms of a criterion-referenced test, at least three alternative approaches are possible. The first is to simply specify a passing score on a test without any reference to some notion of true score. (See Huynh, 1976; Subkoviak, 1976; Wilcox, 1979a.) The second is to take the view that examinees either know or do not know the answer to an item on a test, and the goal is to determine which of the n items an examinee really knows. The third view is that the items represent a larger domain of items, and the goal is to determine the proportion of items in the item pool that the examinee knows. The latter two views are discussed below.

4. Latent State Models

Latent state models (also known as latent structure or latent class models) have existed for some time (e.g., Lazarsfeld & Henry, 1968; Lazarsfeld, 1950). One of the original applications was measuring attitudes (Stouffer, 1950), but only situations involving aptitude and achievement are considered here. Also there are continuous latent structure models that are similar to latent trait models, but only discrete models are discussed.

A basic premise in latent state models is that in terms of a specific item, examinees can be described as belonging to one of

finitely many states. The relative merits of this view are discussed in a more general context by Hilke et al. (1977), Scandura (1971, 1973), and Spada (1977).

The simplest case is where examinees are said to either know or not know the correct response. The obvious problem is that under conventional situations, an examinee's response might not reflect his/her true state. For example, a testee might choose the correct response on a multiple-choice test item without knowing what the correct response actually is. Latent state models make assumptions about the way examinees behave when responding to an item, or they make assumptions about the way items are related to one another (for example, it might be assumed they are hierarchically related), or they assume that examinees respond to the same items on two different occasions in time. Although very general models are available, no one model will be appropriate for every item on every test. An investigator must make a decision about which latent state model is most appropriate and most convenient in a given situation. Once test scores are available, the chosen model can be checked in various ways. For multiple-choice items, it now appears that one of two models will frequently fit most or all of the items on a test (Wilcox, 1982b). If future investigations support this result, it may now be possible to apply latent state models in a relatively straightforward manner.

The purpose of this section is to review general theoretical results on latent state models that are based on one of the three assumptions mentioned above.

Test-Retest Models

As a simple illustration of how latent state models work, suppose an item is administered to a random sample of N examinees on two separate occasions in time. Let ζ be the proportion of examinees in the population of examiness who know the answer, and let β be the probability of correctly guessing the answer when the examinee does not know. In other words, for a randomly sample examinee

$$\beta = \Pr(\text{correct response} \mid \text{examinee does not know}).$$

Let a 1 indicate a correct response, and a 0 an incorrect response to an item. If p_{ij} is the probability of the response pattern ij on the two occasions (i=0, 1; j=0, 1), if no learning takes place between the two administrations, and if the event of correctly guessing is independent on the two testings, the probability of a correct-correct response pattern for a randomly sampled examinee is

$$p_{11} = \zeta + (1 - \zeta)\beta^2 . \tag{4.1}$$

For the remaining three response patterns, it follows that

$$p_{10} = p_{01} = (1 - \zeta)(1 - \beta)\beta \tag{4.2}$$

and

$$p_{00} = (1 - \zeta)(1 - \beta)^2 . \tag{4.3}$$

The p_{ij}'s are not known, but they can be estimated with x_{ij}/N where x_{ij} is the number of examinees who get the response pattern ij. It follows that

$$\beta = 1 - \frac{p_{00}}{p_{10} + p_{00}} . \tag{4.4}$$

Thus, the unknown latent quantity β can be estimated by replacing the p_{ij}'s with x_{ij}/N. Note that the model implies that $p_{10} = p_{01}$ which can be tested (McNemar, 1947). Results on the power of McNemar's test are given by Wilcox (1977a). Also, note that with a large enough sample the model will probably be rejected, but it may be that p_{10} and p_{01} are nearly the same in value.

If $\hat{\beta}$ is the estimate of β using equation (4.4), ζ can be estimated by replacing β with $\hat{\beta}$ in equation (4.1), replacing p_{11} with x_{11}/N, and solving for ζ. Some properties of this estimation procedure are given by Wilcox (1977a). For example, it is shown that if p is the common value of $p_{10} = p_{01}$ under the assumption the model holds, $(x_{10} + x_{01})/N$ is an unbiased, efficient, maximum likelihood estimate of p.

A related and slightly more general model was proposed by Brownless and Keats (1958). In addition to the latent parameters ζ and β, the model includes the proportion of examinees who learn the item between the two administrations, and the proportion of examinees who repeat the same response from memory on the second testing. Not all of the parameters in the Brownless and Keats model can be estimated, but ζ and β can again be determined. For a similar model, see Marks and Noll (1967).

The Brownless and Keats model appears to be one of the earliest attempts to go beyond the simple knowledge or random guessing model that is frequently adopted. Unfortunately, for practical purposes, the models just described are not convenient because they require two administrations of an item.

Models Based on Items That Are Assumed To Be Related in Some Particular Fashion

This section reviews models where items are assumed to be related in a particular fashion. Two situations have been examined in the literature. The first is based on the assumption that two or more items are hierarchically related, and the second is that items are equivalent. Two items are defined to be equivalent if all examinees know both items

or neither one. Of course models for hierarchically related items contain models for equivalent items as a special case. Consider two equivalent items and let ζ be the proportion of examiness who know both. Let p_{ij} be the probability of the response pattern ij on two equivalent items. If β_1 is the probability of correctly guessing the response to the first item when the randomly sampled examinee does not know, and if β_2 is the corresponding probability on the second item, and if local independence holds (i.e., given an examinee's latent state, the responses are independent) then

$$p_{11} = \zeta + (1 - \zeta)\beta_1\beta_2$$

$$p_{10} = (1 - \zeta)\beta_1(1 - \beta_2)$$

$$p_{01} = (1 - \zeta)\beta_2(1 - \beta_1)$$

$$p_{00} = (1 - \zeta)(1 - \beta_1)(1 - \beta_2) .$$

Solving for ζ, β_1, and β_2 yields

$$\beta_2 = \frac{p_{10}}{p_{10} + p_{00}}$$

$$\beta_1 = \frac{p_{10}}{p_{01} + p_{00}}$$

and

$$\zeta = 1 - (p_{01} + p_{00})(p_{10} + p_{00})/p_{00} .$$

Again, the p_{ij}'s can be estimated in the usual manner which yields an estimate of ζ, β_1, and β_2 (Wilcox, 1977b).

Multiple-choice items are the most obvious examples where errors at the item level (guessing) need to be considered. However, even when completion items are used, it may be necessary to measure and correct errors at the item level (e.g., Harris & Pearlman, 1978; Macready & Dayton, 1977). This time though, the quantity of interest is

$$\alpha = \Pr(\text{incorrect response} | \text{examinee knows}) ,$$

and in the simplest case it is assumed that $\beta = 0$. Again ζ and α can be related to the p_{ij}'s. In particular,

$$p_{11} = (1 - \alpha_1)(1 - \alpha_2)\zeta$$

$$p_{10} = (1 - \alpha_1)\alpha_2 \zeta$$

$$p_{01} = \zeta\alpha_1(1 - \alpha_2)$$

$$p_{00} = \zeta\alpha_1\alpha_2 + (1 - \zeta) .$$

Thus,

$$\alpha_1 = p_{01}/(p_{01} + p_{11})$$

$$\alpha_2 = 1 - p_{11}/(p_{10} + p_{11})$$

and

$$\zeta = (p_{01} + p_{11})(p_{10} + p_{11})/p_{11} .$$

Replacing the p_{ij}'s with their ususal unbiased estimate yields an estimate of ζ, α_1, and α_2. For some related models and results see Knapp (1977), Harris and Pearlman (1978), and Harris et al. (1980).

If three or more equivalent items are available, it is possible to estimate both β and α using the procedure outlined by Goodman (1979), or using the scoring method as in Macready and Dayton (1977). These two estimation procedures rely on iterative techniques that approximate the maximum likelihood estimates of the parameters in the model. In practice, these techniques seem to converge very rapidly, and so sometimes they could even be applied when computer facilities are not available (cf. Kale, 1962). However, models can become quite complex necessitating computer facilities.

How can the assumption that two or more items are equivalent be empirically checked? One way is to apply a goodness-of-fit test to the resulting latent structure model as is illustrated by Macready and Dayton (1977). (For some recent results and comments on using goodness-of-fit tests, see Smith et al., 1981; Koehler & Larntz, 1978; Chapman, 1976.) However, this approach is useless in the case of only two items (unless it is assumed that $\beta_1 = \beta_2$ and $\alpha = 0$) because there are then three latent parameters and only four possible response patterns resulting in zero degrees of freedom.

An alternative approach was suggested by Hartke (1978) that is based on latent partition analysis, and an index proposed by Baker and Hubert (1977) might be useful in this endeavor as well. If multiple-choice test items are being used, and if the test is administered according to an answer-until-correct scoring procedure (which is described below), certain equalities are implied when items are equivalent, and these equalities can be tested (Wilcox, 1981d). Some additional possibilities are mentioned by Wilcox (1982f).

Hierarchically Related Items or Guttman Scales

Latent structure models based on the assumption that items are hierarchically related or that the possible latent states form a Guttman scale, include as a special case the notion of equivalent items. In terms of equivalent items, examinees are described as being in one of two states; they know both items or neither one. For two hierarchically related items, a third state is included, namely knowing the second item but not the first. Again, in certain special cases, the proportion of examinees in each of the latent classes can be estimated using simple (closed form) equations. Very general models are also available where estimates are obtained via iterative techniques (Dayton & Macready, 1980, 1976).

As a simple illustration, consider two items and let τ_1 be the proportion of examinees who know the second but not the first. If the guessing rate is the same on the two items, i.e., $\beta_1 = \beta_2 = \beta$, say, then

$$p_{11} = \zeta + \zeta_1\beta + (1 - \zeta - \zeta_1)\beta^2$$

$$p_{10} = \zeta_1(1 - \beta) + (1 - \zeta - \zeta_1)\beta(1 - \beta)$$

$$p_{01} = (1 - \zeta - \zeta_1)\beta(1 - \beta)$$

$$p_{00} = (1 - \zeta - \zeta_1)\beta^2 .$$

It follows that

$$\beta = p_{01}/(p_{01} + p_{00})$$

$$\zeta_1 = (p_{10} - p_{01})/(1 - \beta)$$

$$\zeta = 1 - \beta(1 - \beta)^{-1}p_{01} - \zeta_1 .$$

and so maximum likelihood estimates are easily obtained (Wilcox, 1980a). This model is restrictive in the sense that $\beta_1 = \beta_2$ might be untenable, but much more general models are available which allow $\beta_1 \neq \beta_2$ (Dayton & Macready, 1976, 1980).

Verifying Hierarchies

Interest in learning hierarchies has been with us for some time (e.g., Gagné & Paradise, 1961; Gagné, 1968; Cox & Grahman, 1968) but here attention is focused only on the role latent structure models play in verifying hierarchies. Apparently the first method of examining whether two items are hierarchically related was proposed by White and Clark (1973). The procedure is based on the assumption that for each of the two items being investigated, an equivalent item is available. The probability of the various response patterns can be written in terms of the relevant latent parameters which yields a test of whether the items are hierarchically related. Although White and Clark (1973) were explicitly interested in determining whether two items are hierarchically

related, technically they were not the first to formulate a model that could be used for this purpose. In particular, Proctor (1970) proposed a latent structure model where the latent states of examinees are assumed to form a Guttman scale. A goodness-of-fit test could be used to check whether it is reasonable to assume items are hierarchically related. Today Proctor's model would presumably be replaced by ones proposed by Dayton and Macready (1976, 1980), and again a goodness-of-fit test could be used. However, as was the case for equivalent items, there are situations where this is inappropriate. Again the problem is that there are as many latent parameters as there are degrees of freedom.

A third method is based on an answer-until-correct scoring procedure. If two items are hierarchically related, certain equalities should hold which, for convenience, are described in a later section of the paper.

Some Concluding Remarks on Latent State Models for Equivalent and Hierarchically Related Items

Clearly there are situations where the notion of equivalent or hierarchically related items is too restrictive. This point was raised by Molenaar (1981), and the author would certainly concur. However, there are situations involving real data where the notion of equivalent items seems to be useful (Macready & Dayton, 1977; Harris & Pearlman, 1978). More recently, Harris et al. (1980) applied an equivalent item model to real data collected in school settings. This was done every week over a period of many weeks. All indications were that the test results provided valuable and valid information. Moreover, these models allow $\alpha > 0$, while the models described in the next section assume $\alpha = 0$.

Methods of estimating the parameters in latent structure models were already mentioned, and typically these are used. For some related results see Harris et al. (1980), Rao (1973), Wilcox (1977a, 1977b, 1980a, 1980b), Haberman (1977), Werts et al. (1973), and van der Linden (1981).

For some related general results and comments on latent structure models, see McHugh (1956), Keesling (1974), Bergan et al. (1980), Reulecke (1977) Lazarsfeld and Henry (1968), Gibson (1959, 1962), Goodman (1974), Green (1951), and Gilula (1979). For additional comments on how latent structure models relate to latent trait models, see van der Linden (1978). For an approach to measurement problems that is somewhat related to the discussion in this subsection, see Cliff (1977) and Harnisch and Linn (1981).

Models Based on Assumptions About How Examinees Behave When Taking Multiple-Choice Test Items

Despite the very general nature of the model discussed by Dayton and Macready (1980), and some recent related results reported by Macready and Dayton (1980a) and Bergan et al. (1980), there remains the practical problem of initially determining how items relate to one another so that a particular latent structure model can be tried out on observed test scores. Another potential problem is that the items on a particular n-item test might not be consistent with any particular form of the model. For practical purposes it would be convenient to have a model that could be used to measure the effects of guessing without assuming that items are related in any particular fashion. It would also be helpful if the model were easy to use, i.e., it could be used in a classroom with minimal effort. A third desirable property, one related to the first, would

be the ability to easily fit a simple model to all the items on an arbitrarily chosen n-item test. This last goal was reached in Wilcox (1982b). Before indicating how this was done, some earlier results will be given first.

Suppose multiple-choice test items are scored according to an answer-until-correct (AUC) scoring procedure. This means that examinees choose an alternative, and they are told immediately whether they are correct. One way to accomplish this is to have examinees erase a shield on a specially designed answer sheet which is available commercially. Underneath the shield is an indication of whether the examinee is correct. If incorrect, the testee chooses another response, and this process continues until the correct alternative is identified.

Unlike other latent structure models, Wilcox (1981c) makes certain assumptions about how examinees behave when responding to a multiple-choice item, namely, that examinees eliminate as many distractors as they can (through partial information) and then guess at random from among the alternatives that remain. This assumption is not new (e.g., Horst, 1933), but it was not previously used in conjunction with latent state models. Undoubtedly this assumption is an over simplification of reality, but it has proven to be consistent with most of the items studied by Wilcox (1982a, 1982b, in press a).

For a randomly sampled examinee responding to a particular item, let ζ again be the probability the examinee knows the answer, and let ζ_i ($i=1, \ldots, t-2$) be the probability the examinee can eliminate i distractors if he/she does not know, where t is the number of alternatives. If p_i is the probability that a randomly selected examinee gets the correct response on the ith attempt of an item, then

$$p_1 = \varsigma + \sum_{i=0}^{t-2} \varsigma_i/(t - i) \qquad (4.6)$$

and

$$p_i = \sum_{j=0}^{t-i} \varsigma_j/(t - j) \quad (i=2, \ldots, t) \qquad (4.7)$$

It follows that $\varsigma = p_1 - p_2$. Thus, if in a sample of N examinees, x_i are correct on the ith attempt, then

$$(x_1 - x_2)/N \qquad (4.8)$$

is an estimate of ς. The model implies that

$$p_1 \geq p_2 \geq \cdots \geq p_t \qquad (4.9)$$

and this can be tested (Robertson, 1978). Empirical investigations (Wilcox, 1982a, 1982b) suggest that (4.9) will frequently hold.

Equation (4.9) rules out the misinformation model proposed in Wilcox (1982b), but it is difficult to say whether testing (4.9) gives a strong indication of whether the model holds. Perhaps some other model could be derived that explains existing data (e.g., Hutchinson, 1982). In addition, the random guessing component of the model is undoubtedly untrue (i.e., examinees guessing at random once they eliminate as many distractors as possible). However, an empirical investigation into an implication of the random guessing component of the model suggested that the model gives a tolerable approximation of reality (Wilcox, in press a). When this investigation was conducted, it was thought that a generalization of the AUC model would be needed that takes into account the order in which distractors are chosen. So far, though, it seems that the simpler model described above will suffice.

The latent structure model just described implies that equation (4.9) must hold for the population of examinees. In a few instances this assumption appears to be unreasonable, and the question arises as to how these results might be explained. The solution proposed by Wilcox (1982b) is that some of the examinees have misinformation relative to the question being asked. This appeared to be a reasonable speculation based on the way the questions were phrased, and so a modification of the answer-until-correct scoring procedure was proposed. For example, one of these items dealt with the weight of iron after being heated. The examinees (who were approximately 14 years old) were told that when heated, iron expands. They were also told the weight of the iron before it was heated. They were then asked what the weight of the iron would be when red hot. Three of the alternatives were weights that were higher than the weight at room temperature. Thus, it seems reasonable that some examinees might believe that iron is heavier because it expands, and they would therefore choose among the three alternatives consistent with this belief.

In contrast to earlier models, it was decided to derive a latent structure model where examinees belong to one of three latent states rather than only two, namely, they know the answer, they have misinformation as just described, or they are in complete ignorance and guess at random. The resulting model gave a good fit to the data, and a similar model was derived for the other item that did not fit the original answer-until-correct model described above. The point that is particularly interesting is that observed responses to all 30 items on the test could

be explained with models that are very easy to use.

Despite the advantages of this model, there may be situations where certain features are objectionable. For example, the model assumes that an item has at least one effective distractor for those examinees who do not know. Put another way, it is assumed that no distinction is made between examinees who know, and those who can eliminate all of the distractors. For practical purposes, the seriousness of this problem is not known. Another feature is that it assumes α = Pr(incorrect response|examinee knows) = 0. Again the seriousness of this restriction is not well understood.

Some Miscellaneous Models

In addition to the models described so far, three slightly related models have been proposed by Reulecke (1977). The first, which Reulecke calls the Poisson-binomial model, assumes that examinees are responding to n equivalent items. For examinees who know, it is assumed that they give an incorrect response to x items with probability $h^x \exp(-h)/x!$, the Poisson density, where h is an unknown parameter. For examinees who do not know, it is assumed that $\beta = .5$. His second model replaces the assumption that $\beta = .5$ with the assumption that guessing x items given the examinee does not know is $u^{n-x}\exp(-x)/(n-x)$ where u is an unknown parameter. The third model is the same as the last except that an additional latent state is included, namely, that some examinees guess at random.

An alternative approach to measuring misinformation was proposed by Duncan (1974). For a particular n item test, let δ_1 be the number of items an examinee knows, and let δ_2 be the number of items for which the examinee has misinformation. If every item has t alternatives, Duncan assumes that guessing is at random, and that the probability of getting x items correct is

$$f(x|\delta_1,\delta_2) = \binom{n-\delta_1-\delta_2}{x-\delta_1} t^{\delta_1-x}\left[\frac{t-1}{t}\right]^{n-\delta_1-x}$$

Both Bayesian and empirical Bayesian estimates of δ_1 are discussed.

5. Applications of Latent Class Models, and the Need To Correct For Guessing

Latent class models can now be used to analyze items, analyze n-item tests, and they can be used when an item sampling model is deemed appropriate. This section outlines the procedures that are available. The main advantages of these procedures are that they provide ways of dealing with guessing that are not possible with other models. But why worry about guessing? Perhaps guessing will have little effect on the purpose of a test. Of course answering this question is crucial in order to motivate the procedures described here, and so a few comments will be made along these lines.

Let ω be the proportion of items in a domain of items that an examinee knows, and suppose the goal is to determine whether $\omega \geq \omega_0$ for some predetermined ω_0. This problem has received considerable attention in recent years as evidenced by the 1980 special issue of Applied Psycholog-

ical Measurement. Suppose $\omega_0 = .8$, and that it is desired to choose n, the test length so that the probability of correctly determining whether $\omega > \omega_0$ is at least .9 whenever $\omega \geq .9$ or $\omega \leq .7$. From Wilcox (1980b), n=29 items are required. Van den Brink and Koele (1980) pointed out that even if random guessing can be assumed, about five or six times as many items are needed to ensure the same level of accuracy as when there is no guessing. Wilcox (1980b) noted that random guessing can not be assumed in which case over 2,600 items are needed.

As another illustration, Ashler (1979) observed that guessing can seriously affect the estimate of the biserial correlation.

A third reason to be concerned about guessing is that it might be important to determine how many items on a test an examinee knows, or even which items are known and which are not. Surely, this is sometimes important when measuring achievement, but guessing can seriously affect the results. An illustration with real data is given in Wilcox (1982d).

Finally, most solutions to measurement problems ignore guessing, or assume guessing is at random. Perhaps one of these assumptions will give reasonable results in some situations, but all indications are that this is not always the case. In fact guessing seems to be more serious than might at first be expected, and so it seems that there might be few measurement problems where guessing can be ignored. It might appear that certain latent trait models handle guessing, but this is not necessarily the case because the type of guessing examined in latent class models is different from the type of guessing in latent trait models. This point is elaborated in Section 5.

The way latent class models are applied will depend in part on whether an item sampling view is believed to be appropriate, whether operational versions of the test are to be based on conventional scoring or AUC scoring, or whether items can be assumed to be related in a particular fashion. If conventional scoring is to be used, then preliminary investigations of a test might be made via AUC scoring to determine which items are particularly affected by guessing, and to measure the overall accuracy of a specific n-item test. Methods for solving these problems are outlined below.

Analyzing an n-Item Test

Consider an n-item test, and suppose the goal is to determine how many items an examinee knows. Further suppose that it is decided an examinee knows if and only if the correct response is given. How accurate is the test for the typical (randomly sampled) examinee?

Let τ_i be the probability of making a correct decision about whether an examinee knows or does not know the ith item when a conventional scoring procedure is used. The parameter τ_i is easily estimated under an answer-until-correct scoring procedure; it is one minus the probability of a correct response on the second attempt (Wilcox, 1981c). A natural way to characterize an n-item test is to use

$$\tau_s = \sum \tau_i$$

the expected number of correct decisions for a randomly sampled examinee who takes the test.

In some cases some additional related information is useful. Suppose,

for example, there are n = 10 items, and τ_s is estimated to be 7. That is, the expected number of items for which a correct decision is made about what an examinee knows is estimated to be 7. To get a better indication of how well the test is performing it would be useful to also know the likelihood of say at least 8 correct decisions among the n = 10 items. Knowing τ_s does not yield much information about this value.

More generally, let ρ_k be the probability of making at least k correct decisions among the n items about whether a typical examinee knows. Certainly ρ_k is a useful measure of how well a test indicates what a typical examinee knows. If τ_s or ρ_k is judged to be too small, the test needs to be modified in some way. For example, the number of distractors might be increased, or perhaps the existing distractors might be improved.

The parameter ρ_k can be expressed symbolically, and more precisely, in the following manner. Suppose it is decided that a testee knows the answer to an item if and only if the correct response is given on the first attempt of the item. For a randomly sampled examinee, let $y_i = 1$ if a correct decision is made about the examinee's latent state on the ith item; otherwise $y_i = 0$. Then

$$\rho_k = Pr(\sum y_i \geq k).$$

Wilcox (1981 b, 1982f) refers to ρ_k as the k out of n reliability of a test.

In classical test theory, the reliability of a test can be estimated if two parallel forms exist. Of course, no two tests are ever exactly parallel, and so bounds on the reliability are used instead. The best known bounds are the Kuder-Richardson formulae. These bounds are expressed in terms of unknown population parameters such as the difficulty level of the items on the test, and variance of the test scores. Although these parameters are not known, they can be estimated. A similar situation occurs in terms of estimating ρ_k. If it can be assumed that y_i is independent of y_j, $i \neq j$, ρ_k could be estimated (Wilcox, 1982c). However, there may be cases where this independence does not hold in which case there is no method of estimating ρ_k. However, both upper and lower bounds on ρ_k are available, and these can be estimated (Wilcox, 1982f, 1981c). Even if y_i and y_j are independent, estimating ρ_k can be a computationally tedious process when n is large, and so again these bounds might be useful.

In the event y_i and y_j are independent for all i and j, it is also possible to make inferences about whether ρ_k is large or small (Wilcox, 1982c). Unfortunately, there is currently no empirical procedure for determining when this independence might hold, and so some caution should be exercised.

More recently, Wilcox (in press b) proposed an approximation of ρ_k that appears to work well when n is small, say $n \leq 5$. For larger values of n the Bonferroni inequality can be applied as indicated by Wilcox.

What To Do When τ_s or ρ_k is Too Small

If the estimate of τ_s or ρ_k is judged to be too small, two general approaches are available. First, identify which items are seriously affected by guessing, and either increase the number of distractors, or attempt to improve the ones that are being used. The second approach is to use a scoring procedure based on an AUC test proposed by Wilcox (1982e). However, the effectiveness of Wilcox's scoring procedure is not known when the number of examinees is small. An investigation into this problem is underway.

If the first approach is selected, two measures are available for deciding whether distractors for an item are working well. The first is to use some Schur function (see Marshall and Olkin, 1979), such as

$$H(p_2, \ldots, p_t) = -\sum\left(\frac{p_i}{1 - p_1}\right)\ln\left(\frac{p_i}{1 - p_1}\right)$$

the entropy function which measures how "far away" guessing is from being random. H is also known as Shannon's measure of information or diversity. If guessing is at random, in which case the distractors have achieved their maximum effectiveness,

$$p_2 = p_3 = \ldots = p_t$$

and H attains its maximum value. Its minimum value occurs when $p_1 = 1 - p_2$ and $p_3 = p_4 = \ldots = p_t = 0$, in which case guessing is as far away from being random as it can be.

Wilcox (1981c) proposed another measure of how well the distractors are performing. Labeled Δ, it is just the difference between the maximum possible value of τ (for fixed ζ) and the actual value of τ. An illustration of the Δ measure is given in Wilcox (1982b).

The entropy function measures the extent to which p_2, p_3, ..., p_t are unequal; the closer the distractors are to being equal, the closer is the item to the ideal situation where guessing is at random. H can be estimated by replacing the p_i's with x_i/N. This yields a maximum likelihood estimate of H, say $\hat{H}$, but the exact distribution of $\hat{H}$ is complex and cumbersome to work with, and an asymptotic approximation of the distribution of $\hat{H}$ tends to be unsatisfactory unless N is very large (Bowman et al., 1971). Accordingly, it might be convenient to have some other index that measures the extent to which p_2, ..., p_t are unequal. It turns out that a whole family of functions exists that have properties similar to H (Marshall & Olkin, 1979). One function that seems especially convenient is Simpson's measure of diversity (Simpson, 1949). For the situation at hand it is given by

$$S = \sum_{i=2}^{t} [p_i/(1 - p_1)]^2 .$$

Note that random guessing can be tested by testing $p_2 = p_3 = \cdots p_t$ (see Smith et al., 1981, Wilcox, 1982e). But if the null hypothesis is rejected, the real question is how far away the item is from random guessing, and the measures S and H answer this question.

Alam and Mitra (1981) report some results on the distribution of

$$\sum_{i=1}^{t} x_i^2$$

which might be used to make inferences about S, but there is an error in their results. Alam (1981) confirms the error, and a correction is in preparation.

Testing Whether Items are Equivalent or Hierarchically Related

The same model used to make inferences about ρ_k can also be used to test whether items are equivalent or hierarchically related. The procedure can be briefly outlined as follows. For a randomly sampled examinee responding to a pair of specific test items, let τ_{ij} be the probability of being able to eliminate i distractors from the first item and j distractors from the second. The proportion of examinees who know both items corresponds to $\tau_{t-1,t-1}$ where t is still the number of distractors. If p_{km} is the probability of a correct response on the kth and mth attempt of the two items respectively, then under certain mild independence assumptions,

$$p_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t-m} \tau_{ij}/[(t-i)(t-j)] .$$

If the two items are hierarchically related, then some of the τ_{ij}'s must equal zero, which in turn means that some of the p_{km}'s will be equal to one another. An illustration is given in Wilcox (1982f). These equalities can be tested in numerous ways yielding an empirical check on whether the items are hierarchically related.

Correcting for Type II Guessing

All of the applications that have been described are based on what Wilcox (1981c) calls Type I guessing. This just means that guessing is defined in terms of a randomly sampled examinee responding to a

randomly sampled item. That is, an examinee's guessing ability is the probability of giving a correct response to a typical test item that he/she does not know. The situation is similar to the item sampling models described earlier, except that guessing is taken into account. Rather than estimating ξ, an examinee's percent correct true score, the goal is to estimate ω, the proportion of items in the item pool that the examinee knows.

It is a simple matter to adjust latent structure models developed under Type I guessing to the problem of estimating ω (e.g., Wilcox, 1979b, 1981c, 1982a). Consider, for example, an answer-until-correct test. If, for a specific examinee, ω_i is the probability that he/she can eliminate i distractors from a randomly chosen item, then the probability of getting an item correct on the first attempt, is

$$q_1 = \omega + \sum_{i=0}^{t-2} \omega_i/(t - i)$$

and the probability of a correct response on the second attempt is

$$q_2 = \sum_{i=0}^{t-2} \omega_i/(t - i)$$

so

$$\omega = q_1 - q_2 .$$

Thus, if on an n-item test there are z_i items for which the examinee is correct on his/her ith attempt, the estimate of ω is simply

$$\hat{\omega} = (z_1 - z_2)/n .$$

Indeed, all of the results under Type I guessing are also available under Type II guessing.

Should interest be directed toward determining which (or how many) of the n items on a test an examinee knows, or toward estimating ω, or both? Macready and Dayton (1977, 1980) argue that at least in some situations, the former goal should be sought, and that perhaps formulating the goal of a test in terms of ω should be avoided. It would seem that the solution to this problem will depend on what exactly an investigator wants to determine, and of course this will vary from situation to situation.

An advantage of estimating ω with an answer-until-correct scoring procedure is that it can substantially reduce the problems noted by van den Brink and Koele (1980), and Wilcox (1980b) when trying to determine whether ω is above or below some known constant. This is one of the problems mentioned at the beginning of this section. In situations where an answer-until-correct scoring procedure can be used, there are now two related solutions that might be adopted (Wilcox, 1982g, (1982d). The former approach is particularly well suited for computerized testing where a sequential scoring rule can be used.

Strong True Score Models

As previously indicated, the Type II guessing model under the answer-until-correct procedure implies that ω, the proportion of items an examinee knows, is equal to $q_1 - q_2$, where q_i is the probability of a correct response on the ith attempt of a randomly selected item. Under a conventional scoring procedure where an examinee gets only one attempt at an item, q_1 is the examinee's percent correct true score. If for a population of examinees, the distribution of q_1 could be determined,

many practical measurement problems could be solved (Lord, 1965; Lord & Novick, 1968, ch. 23; Wilcox, 1981a). The most frequently used approach when estimating this distribution, say $g(q_1)$, is to assume that

$$g(q_1) = \frac{\Gamma(r + s)}{\Gamma(r)\ (s)} q_1^{r-1}(1 - q_1)^{s-1} \tag{5.1}$$

the beta density with parameters $r > 0$, and $s > 0$, and where Γ is the gamma function. Empirical studies cited by Wilcox (1981a) indicate that (5.1) will frequently give good results when addressing various measurement problems.

Is it possible to develop a similar strong true score model that takes into account the guessing ability of the examinees? Wilcox (1981a) summarized results on several models that have been proposed, and so they will not be discussed here. The important point is that all of the strong true score models reviewed by Wilcox (1981a) now appear to be totally inadequate for both theoretical and empirical reasons. Some of these models were based on the assumption that guessing is at random, but recent empirical investigations indicate that this is highly unsatisfactory (Wilcox, 1982a, 1982b). See also Bliss (1980) and Cross and Frary (1977). Other models were based on a multivariate analog of the beta-binomial distribution (the Dirichlet-multinomial) which allowed β to vary over the population of examinees. This model implies that ω and β are independent (Wilcox, 1981b) but this appears to be an unsatisfactory assumption because the model gives a very poor fit to data.

Coombs et al. (1956) suggested that an examinee's guessing ability increases with the proportion of items he/she knows. Wilcox (1982a) proposed a strong true score model based on this assumption and an answer-until-correct scoring procedure under Type II guessing. Among the

several models that were considered, this was the only model that gave a reasonable fit to the data. A more recent empirical study got similar results (Wilcox, 1982b).

The model assumes (5.1) holds, and as already mentioned, this frequently gives good results with real data. Let $\gamma = q_2/(1 - q_1)$. The model also assumes that γ can be written as an increasing function of q_1 which is given by

$$\gamma(q_1) = c\int_0^{q_1} \frac{\Gamma(r_1 + s_1)}{\Gamma(r_1)\Gamma(s_1)} u^{r_1-1}(1 - u)^{s_1-1} du + (t - 1)^{-1}$$

where c, $r_1 > 0$, and $s_1 > 0$ are unknown parameters that are estimated from observed test scores. (The subscripts on r and s are used to distinguish them from the parameters r and s used earlier.) A method of estimating c, r_1, and s_1 is described by Wilcox (1982a).

This model can be used to solve many measurement problems that were previously impossible to solve. For example, suppose a conventional test is administered, and it is desired to correct for guessing without assuming guessing is at random. If the function $\gamma(q_1)$ has been previously estimated, then $\omega = q_1 - \gamma(q_1)$. If γ is arbitrarily set equal to $(t - 1)^{-1}$, the usual correction for guessing formula score results.

It should be mentioned that while it is possible to correct for guessing under the answer-until-correct procedure, alternative scoring rules might be preferred (Brown, 1965; Dalrymple-Alford, 1970). These scoring formulae do not estimate ω, but instead give an examinee credit for having partial information. Whether this is desirable will depend on the examiner's goal. Of course, several other scoring procedures have been proposed, some of which are discussed by Frary (1980). The important point is that none of these rules yields an estimate of ω. The same is true of the procedure proposed by Gibbons, Olkin, and Sobel (1979), and the rule suggested by Austin (1981). Note that Austin's procedure is the same as one proposed by Arnold and Arnold (1970) which is discussed by Frary (1980).

Additional Applications

Several other applications of latent class models have been examined in the literature which are only mentioned here. These include a tailored testing procedure (Wilcox, in press b) that might be used when computerized testing is feasible. Knapp (1977) discusses a a reliability coefficient that is based on a latent state point of view, and Emrick (1971) describes how these models might be used to determine the passing score of a criterion-referenced test. Emrick's estimation procedure was shown to be incorrect (Wilcox & Harris, 1977), but this problem is easily corrected using one of the estimation procedures already described. A closed form estimate of the parameters in Emrick's model is given by van der Linden (1981).

6. Further Comments on How Latent Class and Latent Trait Models are Related

In the three parameter latent trait model given by equation (2.4), the parameter c is sometimes called a guessing parameter. Hopefully by this point it can be seen that this parameter has nothing to do with the notion of guessing used in latent class models. The parameter c is just $\lim_{\theta \to \infty} p(\theta)$. Thus, c refers to the probability of a correct response to an item for a particular type of examinee, namely, examinees for whom θ is small. For latent class models guessing is defined in terms of a specific item and a population of examinees who do not know, or a specific examinee and a domain of items that he/she does not know--this is different from the population of examinees having θ small. Suppose for exampl

$p(\theta) = \frac{1}{2}$. Using the item sampling interpretation of $p(\theta)$, this means that among all the items having item parameters a, b, and c, the probability of a correct response is ½ for an examinee with ability level θ. But this suggests that the examinee does not know all of these items, in which case some answers will be correct by chance. But how does the parameter c correct this difficulty? The answer is that it doesn't deal with this problem at all.

Some writers have interpreted $p(\theta)$ in (2.4) as the probability of knowing an item which suggests that latent trait models might be related to latent class models, but no simple relationship has been established when errors at the item level exist because the models measure different things. In fact, if this interpretation is used, all estimates of the item parameters in (2.4) break down when multiple-choice items are used. To see this, note that in order to estimate a, b, and c, it would be necessary to determine which items (or how many items) an examinee knows. But what is observed is only which items were answered correctly. In some cases perhaps this is not a serious problem--it seems that more work is needed in this area. Mislevy and Bock (1982), as well as Wainer and Wright (1980) have given some attention to the problem of estimating latent trait parameters in the presence of guessing. However, the model they used for guessing behavior is different from the notion of guessing in latent class models.

To further differentiate the two models, perhaps a more general theoretical description of true score models will help. Sirotnik and Wilcox (1982) point out that certain notions in Torgerson (1958) can be used to describe a model that contains as a special case all the true score models described in this paper. Their developments are briefly summarized here.

Let ψ be some "ability" parameter that characterizes an examinee. For a randomly sampled examinee, let $p_i(\psi)$ be the conditional probability of a correct response to the ith item on an n-item test given that the examinee has ability ψ. Let $p_x(\psi)$ be the conditional probability (given ψ) of x correct responses, and let $g(\psi)$ be the probability density function of ψ for the population of examinees. Then

$$p_i = \int p_i(\psi)g(\psi)d\psi$$

is the probability of a correct response to the ith item for a randomly sampled examinee.

A basic problem is determining what ψ should represent. For a latent class model, the simplest case is for a single examinee and a single item in which case the only two possible values for ψ are 1 (the examinee knows) or 0 (the examinee does not know). Then $g(\psi)$ is the proportion of examinees who know. For the AUC model the possible values of ψ are $0,1,\ldots,t-1$, and $p_i(\psi)=(t-\psi)^{-1}$ for a randomly sampled examinee. Note that for these models, an examinee's ability is defined in terms of a specific item, and this can be used as a basis for defining ability in terms of the number of items known on an n-item test, or the proportion of items known in an item domain. For latent trait models ψ does not indicate what an examinee knows, but rather, it determines the probability of knowing when there are no errors at the item level such as guessing. Another important point is that to say the item parameter c is the same as the guessing parameter in the AUC model is to somehow equate c to $p_i(\psi)$ given that $\psi<t-1$.

For an item sampling model based on a latent class model, ψ is the proportion of items in an item domain that an examinee knows, $0\leq\psi\leq1$, and $p_x = \binom{n}{x}\psi^x(1-\psi)^{n-x}$. In latent trait models, the probability of a correct response to the ith item depends on a, b, and c. Thus, as previously pointed out, for latent trait models, $p_i(\psi)=E_{abc}(\psi)$, where E_{abc}

means expectation with respect to a, b and c. Also

$$p_i = \int\int\int\int p_i(\psi) g(\psi,a,b,c)\, d\psi\, da,\, db,\, dc$$

where $g(\psi, a, b, c)$ is the joint density of ψ, a, b, and c.

7. Concluding Remarks

As was stressed at the beginning of the paper, it is not being argued that the other approaches to measurement (classical test theory, latent trait models, and item sampling models) be abandoned, or that they are intrinsically bad in any sense. It is being argued though that careful examination of the goal of a test should be made before a true score model is chosen. Generally different models give different solutions to the same problem. For example, when determining how many distractors should be used, latent trait models can be applied (Lord, 1980), but the criterion used is different from the one used in latent state models.

Another reason for choosing a model carefully is that some writers have argued that latent trait models do not address many of the measurement problems that are currently of interest (e.g., Baker, 1977). The primary point in this paper is that latent class models give the test constructor ways of examining measurement problems that did not exist a short while ago. By using latent class models in conjunction with other true score models, tests can be analyzed in a more effective manner than ever before.

REFERENCES

Alam, K. Personal communication, 1981.

Alam, K., & Mitra, A. Polarization test for the multinomial distribution. Journal of the American Statistical Association, 1981, 76, 107-109.

Arnold, J. C., & Arnold, P. L. On scoring multiple-choice exams allowing for partial knowledge. The Journal of Experimental Education, 1970, 39, 8-13.

Ashler, D. Biserial estimators in the presence of guessing. Journal of Educational Statistics, 1979, 4, 325-355.

Austin, J. D. Grading distractor-identification tests. Psychometrika, 1981, 46, 129-138.

Baker, F. B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178.

Baker, F. B., & Hubert, L. J. Inference procedures for ordering theory. Journal of Educational Statistics, 1977, 2, 217-233.

Bergan, J. R., Cancelli, A. A., & Luiten, J. W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. Journal of Educational Statistics, 1980, 5, 65-81.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. Novick (Eds.), Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Bowman, K., Hutcheson, K., Odum, E., & Shenton, L. Comments on the distribution of indices of diversity. In G. Patil, E. Pielou, & W. Waters (Eds.), International Symposium on Statistical Ecology (Vol. 3). University Park: Pennsylvania State Press, 1971.

Brown, J. Multiple response evaluation of discrimination. The British Journal of Mathematical and Statistical Psychology, 1965, 18, 125-137.

Brownless, V. T., & Keats, J. A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Chapman, J. W. A comparison of the X^2, -2logR and multinomial probability criteria for significance tests when expected frequencies are small. Journal of the American Statistical Association, 1976, 71, 854-863.

Cliff, N. A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 1977, 42, 375-399.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial information. Educational and Psychological Measurement, 1956, 16, 13-37.

Cox, R. C., & Graham, G. T. The development of a sequentially scaled achievement test. Journal of Educational Measurement, 1966, 3, 147-150.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dalrymple-Alford, E. C. A model for assessing multiple-choice test performance. British Journal of Mathematical and Statistical Psychology, 1970, 23, 199-203.

Dayton, C. M., & Macready, G. B. A probabilisitic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Dayton, C. M., & Macready, G. B. A scaling model with response errors and intrinsically unscalable respondents. Psychometrika, 1980, 45, 343-356.

Diamond, J., & Evans, W. The correction for guessing. Review of Educational Research, 1973, 43, 181-191.

Duncan, G. T. An empricial Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 1974, 69, 50-57.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Frary, R. B. The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. Applied Psychological Measurement, 1980, 4, 79-90.

Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 1961, 75, 1-23.

Gagné, R. M. Learning hierarchies. Educational Psychologist, 1968, 6, 1-9.

Gibbons, J. D., Olkin, I., & Sobel, M. A subset selection technique for scoring items on a multiple choice test. Psychometrika, 1979, 44, 259-270.

Gibson, W. A. Three multivariate models: factor analysis, latent structure analysis, and latent profile analysis. Psychometrika, 1959, 24, 229-252.

Gibson, W. A. Extending latent class solutions to other variables. Psychometrika, 1962, 27, 73-81.

Gilula, Z. Singular value decomposition of probability matrices: Probabilistic aspects of latent dichotomous variables. Biometrika, 1979, 66, 339-344.

Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 1974, 61, 215-231.

Goodman, L. A. On the estimation of parameters in latent structure analysis. Psychometrika, 1979, 44, 123-128.

Green, B. F. A general solution for the latent class model of latent structure analysis. Psychometrika, 1951, 16, 151-166.

Haberman, S. J. Product models for frequency tables involving indirect observation. The Annals of Statistics, 1977, 6, 1124-1147.

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D. R., & Gifford, J. Developments in latent trait theory: models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Harnisch, D. L., & Linn, R. L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18, 133-146.

Harris, C. W., & Pearlman, A. An index for a domain of completion or short answer items. Journal of Educational Statistics, 1978, 3, 285-304.

Harris, C. W., Houang, R. T., Pearlman, A. P., & Barnett, B. Final report submitted to the National Institute of Education. Grant No. NIE-G-78-0085, Project No. 8-0244, 1980.

Hartke, A. R. The use of latent partition analysis to identify homogeneity of an item population. Journal of Educational Measurement, 1978, 15, 43-47.

Hilke, R., Kempf, W. F., & Scandura, J. M. Deterministic and probabilistic theorizing in structural learning. In H. Spada and F. Kempf (Eds.), Structural models of thinking and learning. Bern: Hans Huber, 1977.

Holland, P. W. When are item response models consistent with observed data. Psychometrika, 1981, 46, 79-92.

Horst, P. The difficulty of a multiple choice test item. Journal of Educational Psychology, 1933, 24, 229-232.

Hutchinson, T. P. Some theories of performance in multiple choice tests, and their implications for variants of the task. British Journal of Mathematical and Statistical Psychology, 1982, 35, 71-89.

Huynh, H. On the reliability of decisions in domain-oriented testing. Journal of Educational Measurement, 1976, 13, 253-264.

Kale, B. K. A note on a problem in estimation. Biometrika, 1962, 49, 553-557.

Keesling, J. W. Empirical validation of criterion-referenced measures. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles: Center for the Study of Evaluation, monograph no. 3, 1974.

Knapp, T. R. The reliability of a dichotomous test-item: A "correlationless" approach. Journal of Educational Measurement, 1977, 14, 237-252.

Koehler, K. J., & Larntz, K. An empirical investigation of goodness-of-fit statistics for sparse multinomials. Journal of the American Statistical Association, 1980, 75, 333-342.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. (Eds.), Measurement and Prediction. Princeton: Princeton University Press, 1950.

Lazarsfeld, P. F., & Henry, N. W. Latent structure analysis. New York: Houghton Mifflin, 1968.

Lord, F. M. An approach to mental test theory. Psychometrika, 1959. 24, 283-302.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology, (Vol. 2). San Francisco: Freeman, 1974.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Macready, G. B., & Dayton, C. M. A two-stage conditional estimation procedure for unrestricted latent class models. Journal of Educational Statistics, 1980, 5, 129-156. (a)

Macready, G. B., & Dayton, C. M. The nature and use of state mastery models. Applied Psychological Measurement, 1980, 4, 493-516. (b)

Marks, E., & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.

Marshall, A., & Olkin, I. Inequalities: Theory of majorization and its application. New York: Academic Press, 1979.

McDonald, R. P. The dimensionality of tests and items. The British Journal of Mathematical & Statistical Psychology, 1981, 34, 100-117.

McHugh, R. B. Efficient estimation and local identification in latent class analysis. Psychometrika, 1956, 21, 331-347.

McNemar, Q. Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 1947, 12, 153-157.

Mellenbergh, G. J., & Vijn, P. The Rasch model as a loglinear model. Applied Psychological Measurement, 1981, 5, 369-376.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. Review of Educational Research, 1976, 46, 133-158.

Mislevy, R. J., & Bock, D. R. Biweight estimates of latent ability. Educational and Psychological Measurement, 1982, 42, 725-738.

Molenaar, I. On Wilcox's latent structure model for guessing. British Journal of Mathematical and Statistical Psychology, 1981, 34, in press.

Proctor, C. H. A probabilistic formulation and statistical analysis of Guttman scales. Psychometrika, 1970, 35, 73-78.

Rao, C. R. Linear statistical inference and its application. New York: Wiley, 1973.

Reulecke, W. A. A statistical analysis of deterministic theories. In H. Spada & F. Kempf (Eds.), Structural models of thinking and learning. Bern: Hans Huber, 1977.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

Scandura, J.M. Deterministic theorizing in structural learning. Journal of Structural Learning, 1971, 3, 21-53.

Scandura, J.M. Structural learning: Theory and research. New York: Gordon and Breach, 1973.

Simpson, E. Measurement of diversity. Nature, 1949, 163, 688.

Sirotnik, R., & Wilcox, R. Realizing the potential of latent structure analysis for integrating and differentiating extant true score/ latent ability measurement models. Center for the Study of Evaluation, University of California, Los Angeles, 1982.

Smith, P. J., Rae, D. S., Manderscheid, R. W., & Silbergeld, S. Approximating the moments and distribution of the likelihood ratio statistic for multinomial goodness of fit. Journal of the American Statistical Association, 1981, 76, 737-740.

Spada, H. Logistic models of learning and thought. In H. Spada and F. Kempf (Eds.), Structural models of thinking and learning. Bern: Hans Huber, 1977.

Stouffer, S.A. Measurement and prediction. Princeton: [illegible]ristry Press, 1950.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1980, 33, 104-108.

Van der Linden, W. Forgetting, guessing, and mastery: the Macready and Dayton models revisited and compared with a latent [illegible]t approach. Journal of Educational Statistics, 1978, 3, 305-317.

Van der Linden, W. Estimating the parameters of Emrick's mastery testing model. Applied Psychological Measurement, 1981, 5, to appear.

Wainer, H., Morgan, A., & Gustafasson, J. A review of estimation procedures for the Rasch model with an eye toward longish tests. Journal of Educational Statistics, 1980, 5, 35-64.

Wainer, H., & Wright, B. D. Robust estimation of ability in the Rasch model. Psychometrika, 1980, 45, 373-391.

Weiss, D. J., & Davison, M. L. Review of test theory and methods. Annual Review of Psychology, 1981, 32, 629-658.

Weitzman, R. A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

Werts, C. E., Linn, R. L., & Jöreskog. A congeneric model for platonic true scores. Educational and Psychological Measurement, 1973, 33, 311-318.

White, R. T., & Clark, R. M. A test of inclusion which allows for errors of measurement. Psychometrika, 1973, 38, 77-86.

Wilcox, R.R. New methods for studying stability. In C.W. Harris, A. Pearlman, & R. Wilcox, Achievement test items: Methods of study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977. (a)

Wilcox, R. R. New methods for studying equivalence. In C. W. Harris, A. Pearlman, & R. Wilcox, Achievement Test Items: Methods of Study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977. (b)

Wilcox, R. R. Prediction analysis and the reliability of a mastery test. Educational and Psychological Measurement, 1979, 39, 825-839. (a)

Wilcox, R. R. An alternative interpretation of three stability models. Educational and Psychological Measurement, 1979, 39, 311-316. (b)

Wilcox, R. R. Some results and comments on using latent structure models to measure achievement. Educational and Psychological Measurement, 1980, 40, 645-658. (a)

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement. 1980, 4, 425-446. (b)

Wilcox, R. R. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 1981, 6, 3-32. (a)

Wilcox, R. R. Recent advances in measuring achievement: A response to Molenaar. British Journal of Mathematical and Statistical Psychology, 1981, 34, 229-237. (b)

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414. (c)

Wilcox, R. R. A closed sequential procedure for comparing the binomial distribution to a standard. British Journal of Mathematical and Statistical Psychology, 1981, 34, 238-242.

Wilcox, R.R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1982, 35, 57-70. (a)

Wilcox, R. R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74. (b)

Wilcox, R. R. Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement, 1982, 42, 153-165. (c)

Determining the length of multiple-choice criterion-referenced tests when an answer-until-correct scoring procedure is used. Educational and Psychological Measurement, 1982, 42, 789-794. (d)

Wilcox, R. R. A comment on approximating the X^2 distribution in the equiprobable case. Communications in Statistics -- Simulation and Computation, 1982, 11, 619-623. (e)

Wilcox, R. R. Bounds on the k out of n reliability of a test, and an exact test for hierarchically related items. Applied Psychological Measurement, 1982, 6, 327-336. (f)

Wilcox, R. R. A closed sequential procedure for answer-until-correct tests. Journal of Experimental Education, 1982, 50, 219-222. (g)

Wilcox, R. R. How do examinees behave when taking a multiple-choice test? Applied Psychological Measurement, in press. (a)

Wilcox, R. R. An approximation of the k out of n reliability of a test, and a scoring procedure for determining which items an examinee knows. Psychometrika, in press. (b)

Wilcox, R. R., & Harris, C. W. On Emrick's "An evaluation model for mastery testing." Journal of Educational Measurement, 1977, 14, 215-218.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

STRONG TRUE-SCORE THEORY

Rand R. Wilcox
Department of Psychology
University of Southern California
and
Center for the Study of Evaluation
University of California, Los Angeles

Rand R. Wilcox

Strong True-Score Theory

In mental test theory a general goal is to use observed test scores to make inferences about an unknown parameter θ that represents an examinee's ability in a certain area such as arithmetic reasoning, vocabulary, spatial ability, etc. The parameter θ is frequently called an examinee's true score. There are several types of true scores [3], but because of space restriction the differences among them are not discussed. True score models are just probability models that yield methods for estimating θ or making inferences about the characteristics of a test. The term strong true-score theory was introduced by Lord [2] to make a distinction between "weak" theories that can not be contradicted by data, and "strong" theories where assumptions are made about the distribution of observed test scores. Strictly speaking latent trait models (also known as item response theories) fall within this definition, but the term strong true-score model is usually reserved for models based on the binomial probability function or some related distribution. Apparently this is because the main focus of Lord's paper was a model based on the binomial probability function.

Consider a single examinee responding to n dichotomously scored items. As just indicated the best known strong true-score model assumes that the probability of x correct response is given by

$$f(x|\theta)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x} \qquad (1)$$

In addition to specifying a probability function for x, an examinee's observed score, strong true-score models typically specify a particular family of distributions for θ over the population of examinees. When (1) is assumed the family of beta densities is commonly used where $g(\theta)$, the probability density function of θ, is given by

$$g(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\,\theta^{r-1}(1-\theta)^{s-1}, \qquad (2)$$

and where r, s>0 are unknown parameters. Estimates of the parameters r and s are easily obtained with the method of moments [6] and maximum likelihood estimates are available from [1]. Basically the beta-binomial model falls within the realm of empirical Bayesian techniques, as do most strong true-score models. The beta-binomial model frequently gives a good fit to data, and it provides a solution to many measurement problems [6]. Included are methods of equating tests and methods of estimating test accuracy and reliability.

Several objections have been raised against the beta-binomial model, but from [6] the only objection that seems to have practical importance is that the model ignores guessing. Here a correct guess refers to the event of a correct response to a randomly sampled item that the examinee does not know. For a strong true-score model where a

correct guess is defined in terms of randomly sampled examinees (and where items are fixed), see [12].

Suppose every item has t alternatives, and for a specific examinee let ζ be the probability of knowing a randomly sample item. Morrison and Brockway [4] assumed random guessing in which case

$$\theta = \zeta + t^{-1}(1-\zeta)$$

and the density of θ is

$$g(\theta) = \frac{t}{t-1}\, g\left(\frac{t\theta-1}{t-1}\right), \quad t^{-1} \leq \theta \leq 1.$$

Unfortunately it appears that the random guessing assumption is unsatisfactory. The only model that has given good results is one proposed by Wilcox [8, 9] that is based on an answer-until correct scoring procedure and the assumption that an examinee's guessing ability is a monotonic function of θ. By an answer-until-correct scoring procedure is meant that an examinee chooses responses to a multiple-choice test item until the correct alternative is chosen. These tests are usually administered by having an examinee erase a shield on especially designed answer sheets. Under the shield is a letter indicating whether the correct answer was chosen. If not, another shield is erased, and the process continues until the correct alternative is selected.

Let ζ_i be the probability that an examinee can elimin-

ate i distractors from a randomly sampled item, i=0,1,..., t-1. It is assumed that when an examinee does not know, there is at least one distractor that can not be eliminated through partial information, and so $\varsigma_{t-1} = \varsigma$. It is also assumed that an examinee eliminates as many distractors as possible, and then guesses at random from among the alternatives that remain. For empirical evidence in support of this last assumption, see [11]. If θ_i is the probability of a correct answer on the ith try of a randomly sampled item, then

$$\theta_i = \sum_{j=0}^{t-i} \varsigma_j/(t-j),$$

and so the ς_j's can be estimated. If x_i is the number of items requiring i attempts, it is assumed that the x_i's have a multinomial probability function. It is also assumed that θ_1 has a beta density with parameters r and s, and that

$$E(\frac{\theta_2}{1-\theta_1}|\theta_1)=c\int_0^{\theta_1} h(u)du + t^{-1} \qquad (3)$$

where c is an unknown parameter, and h(u) is also a beta density but with parameters a and b. The model implies that

$$\theta_1 \geq \theta_2 \geq \ldots \geq \theta_t \qquad (4)$$

and so the lower limit for the integral in (3) should be t^{-1}, but this modification has not yet been applied to real data. Equation (3) is based on the assumption that the more

items an examinee knows, the higher the probability will be that an examinee will give a correct guess to an item that is not known. The parameters a, b and c are currently estimated using what is basically the method of moments. The details are too lengthy to report here; the interested reader is referred to [8].

As a final note, there are now extensions of strong true-score models based on closed sequential sampling techniques which might be useful in computerized testing. By closed sequential sampling is meant that items are randomly sampled and administered until some criterion is met. The criterion actually used will depend on the purpose of the test.

Consider, for example, a criterion-referenced test where the goal is to determine whether $\theta \geq \theta_0$ where θ_0 is a known constant. Suppose $\theta \geq \theta_0$ is decided if and only if $x \geq c$, where c (a positive integer) is some known passing score. Given that $\theta \geq \theta_0$ (or that $\theta < \theta_0$), the probability of a correct decision is available immediately (given θ) if the binomial model is assumed. For related results, see [16].

Suppose instead that items are randomly sampled until an examinee gets c items correct or $m=n-c+1$ items wrong. Let x(y) be the number of correct (incorrect) responses when the sampling of items terminates. The joint probability

function of x and y is

$$f(x,y|\theta) = L\,\frac{(x+y-1)!}{x!y!}\,\theta^{x}(1-\theta)^{y},$$

where $x=c$ and $0\leq y\leq m-1$ or where $y=m$ and $0\leq x\leq c-1$, and $L=m$ if $y=m$, otherwise $L=c$. Wilcox [7] showed that the probability of a correct decision under the closed sequential procedure is exactly the same as it is under the binomial model, but the expected number of items is always less. For results on estimating θ under the closed sequential procedure, see [13]. For extensions to the multivariate case, including an application to answer-until-correct tests, see [14, 15].

References

[1] Griffiths, D. A., (1973), Biometrika, 29, 637-648.

[2] Lord, F. M., (1965), Psychometrika, 30, 239-270.

[3] Lord, F. M., & Novick, M. R., (1968), Statistical theories of mental test scores, Addison-Wesley, Reading, Mass. The current classic on mental test theory.

[4] Morrison, D. G., & Brockway, G., (1979), Psychometrika, 44, 427-442.

[5] Wilcox, R. R., (1980), Applied Psychological Measurement, 4, 425-446.

[6] Wilcox, R. R., (1981a), Journal of Educational Statistics, 6, 3-32. A review of the beta-binomial model, with an emphasis on mental test theory.

[7] Wilcox, R. R., (1981), British Journal of Mathematical and Statistical Psychology, 34, 238-242.

[8] Wilcox, R. R., (1982a), British Journal of Mathematical and Statistical Psychology, 35, 57-70. The only item sampling model that has given satisfactory results when dealing with guessing.

[9] Wilcox, R. R., (1982), Journal of Educational Measurement, 19, 67-74.

[10] Wilcox, R. R., (1982), Journal of Experimental Education, 50, 219-222.

[11] Wilcox, R. R., (1983), Applied Psychological Measurement, 8, 239-240.

[12] Wilcox, R. R., (1983), Psychometrika, 48, 211-222.

[13] Wilcox, R. R., (1983), Educational and Psychological Measurement, in press.

[14] Wilcox, R. R., (1982), British Journal of Mathematical and Statistical Psychology, 35, 193-207.

[15] Wilcox, R. R., (1982), Educational and Psychological Measurement, 42, 789-794.

[16] Wilcox, R. R., (1979), Psychometrika, 44, 55-68.

The paper considers the problem of determining whether an examinee has a true score greater than or less than θ_0, an unknown parameter that characterizes a control group.

APPROXIMATING MULTIVARIATE DISTRIBUTIONS

Rand R. Wilcox
Department of Psychology
University of Southern California
and
Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

A simple approximation of a multivariate distribution is suggested that may be useful in certain situations. Comparisons with several other approximations suggest that the new approximation nearly always gives better results. In some cases the improvement is minimal, but for some situations substantially better results are obtained.

Let $X_1,\ldots,X_k$ be k random variables with joint density $f(x_1,\ldots,x_k)$, and let

$$P=Pr(X_1\leq h_1,\ldots,X_k\leq h_k) \qquad (1.1)$$

Of course in many situations (1.1) must be evaluated, but frequently approximations are poor. In some cases P can be evaluated exactly using quadrature techniques, but this can be prohibitively expensive, and the necessary computer programming does not always exist. The goal in this paper is to suggest a simple approximation of P that appears to be useful in various situations, and which appears to compare favorably to some other approximations that have been used in the literature.

The proposed approximation is based on a second order Bahadur approximation of a multinomial distribution. The motivation for this approximation stems from a recent investigation (Wilcox, 1983) which included, among other things, an approximation of $Pr(\sum y_i\geq m)$, where the y_i's are binary random variables. A second order Bahadur approximation proved to be more accurate than expected, and this led to the approximation and comparisons made here. Another motivation for this approximation stems from results reported by McFadden (1955) where it was suggested that a special case of the approximation used here will frequently give good results for k=4.

In section 3 the accuracy of the approximation is investigated by applying it to some distributions where P is known exactly for certain special cases. The results suggest that the approximation nearly always improves upon all four approximations of the multivariate t distribution

proposed by Dunnett and Sobel (1955). In a few instances the improvement is substantial. It also appears to improve upon an approximation of the multivariate normal distribution proposed by Olkin, Sobel, and Tong (1976). Finally, the approximation is compared to some percentage points tabled by Dudewicz and Dalal (1975), and found to give good results in most cases as long as k is not too large. Compared to the Bonferroni inequality, there are again situations where there is considerable improvement.

2. The Approximation

Let $\underset{\sim}{y}=(y_1,\ldots,y_k)$ be a random vector where $y_i=0$ or 1 $(i=1,\ldots,k)$, and let $p(y_1,\ldots,y_k)$ be the corresponding probability function. Bahadur (1961) showed that $p(y_1,\ldots,y_k)$ could be written as

$$p(\underset{\sim}{y}) = p_1(\underset{\sim}{y})g(\underset{\sim}{y})$$

where

$$p_1(\underset{\sim}{y}) = \prod_{i=1}^{k} \alpha_i^{y_i} (1-\alpha_i)^{1-y_i}$$

$$\alpha_i = E(y_i)$$

$$g(\underset{\sim}{y}) = 1 + \sum_{i<j} r_{ij} z_i z_j + \sum_{i<j<m} r_{ijm} z_i z_j z_m + \ldots + r_{12\ldots k} z_1 \ldots z_k$$

$$z_i = (y_i - \alpha_i)/[\alpha_i(1-\alpha_i)]^{\frac{1}{2}}$$

$$r_{ij} = E(z_i z_j)$$

$$r_{ijm} = E(z_i z_j z_m)$$

$$\ldots$$

$$\ldots$$

$$r_{12\ldots n} = E(z_1 z_2 \ldots z_n)$$

An mth order approximation of $p(\underset{\sim}{y})$ is obtained by retaining the first m terms in the expression for $g(\underset{\sim}{y})$. In particular, a second order approximation is

$$p_{\underset{\sim}{i}}(\underset{\sim}{y})\left[1+\sum_{i<j} r_{ij}z_i z_j\right] \tag{2.1}$$

Define

$$y_i = \begin{cases} 1, & \text{if } X_i \leq h_i \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Then an approximation of P is just

$$\left(\prod_{i=1}^{k} \Pr(X_i \leq h_i)\right)\left[1+\sum_{i<j} \frac{\Pr(X_i \leq h_i, X_j \leq h_j)-\Pr(X_i \leq h_i)\Pr(X_j \leq h_j)}{\Pr(X_i \leq h_i)\Pr(X_j \leq h_j)}\right] \tag{2.3}$$

In many practical situations $\Pr(X_i \leq h_i)$ $(i=1,\ldots,k)$ have a common value V, and $\Pr(X_i \leq h_i, X_j \leq h_j)$ have a common value U for all $i \neq j$, in which case (2.3) becomes

$$V^k\left[1+\frac{k^2-k}{2}\left(\frac{U-V^2}{V^2}\right)\right]$$

Bahadur (1961) noted that the approximation (2.1) will be a probability function if $1+\sum_{i<j} r_{ij}z_i z_j \geq 0$, but that otherwise some of its values will be negative. This problem never arose in the cases considered here.

3.1 The Multivariate t Distributions

Suppose the joint probability density function of $X_1,\ldots,X_k$ is multivariate normal with correlation matrix $\{\rho_{ij}\}$, mean vector 0 and common variance σ^2, and nS/σ^2 has a chi-square distribution independent of the X_i's, with ν degrees of freedom. Then the joint density of $T_i=X_i/S$ (i=1.,,,.k) is multivariate t, and the joint pdf (probability density function) is

$$f(t_1,\ldots,t_k)=\frac{A^{\frac{1}{2}}\{\Gamma(\nu+k)/2\}}{n\pi^{k/2}\Gamma(\nu/2)}\left[1+\nu^{-1}\sum_{i,j}a_{ij}t_it_j\right]^{-(\nu+k)/2} \qquad (3.1)$$

where A is the determinant of the positive definite matrix $\{a_{ij}\}=\{\rho_{ij}\}^{-1}$

This distribution arises in ranking and selection (Bechhofer, Dunnett and Sobel, 1954) where the goal is to determine which of k+1 normal distributions has the largest mean. Another application was discussed by Dunnett (1955) where the goal is to compare the mean of k normal distributions to a control. (See, also, Gupta and Sobel, 1958.) Krishnaiah (1965) used the distribution to make multiple comparisons in the multivariate analysis of variance. Properties of this distribution are summarized by Johnson and Kotz (1972) and Gupta (1963).

For k=2 exact expressions for (3.1) are available (Dunnett and Sobel, 1954), but for k>2 approximations must be used except for certain special cases where exact results have been tabulated. An approximation was suggested by John (1961), but unfortunately it is complicated, and some quadrature is required. Four approximations (lower bounds) were proposed

by Dunnett and Sobel (1955). These were

$$1 - \sum \Pr(T_i \geq h_i) \tag{3.2}$$

$$\prod_{i=1}^{k} \Pr(T_i \leq h_i) \tag{3.3}$$

$$\prod_{i=1}^{k/2} \Pr(T_{2i-1} < h_{2i-1}, \Pr(T_{2i} < h_{2i}), \text{ k even} \tag{3.4a}$$

$$\Pr(T_1 < h_1) \prod_{i=1}^{(k-1)/2} \Pr(T_{2i} < h_{2i},\ T_{2i+1} < h_{2i+1}), \text{ k odd} \tag{3.4b}$$

$$\{\Pr(T_1 < h,\ T_2 < h)\}^{k/2} \tag{3.5}$$

The last lower bound assumes $h_1 = h_2 = \ldots = h_k = h$, say, $h \geq 0$, and $\rho_{ij} = \lambda_i \lambda_j$ for some constants λ_i where $0 \leq \lambda_i < 1$ $(i = 1, \ldots, k)$. Expression (3.4) also assumes that $\rho_{ij} = \lambda_i \lambda_j$ and that $h_i \geq 0$ $(i = 1, \ldots, k)$. For $\rho_{ij} = \rho$, Tong (1970) gives the lower bound

$$\{\Pr(T_1 < h,\ T_2 < h) - (\Pr(T_1 < h))^2\}^{k/2}$$

but he shows that this bound is not as sharp as (3.5).

Dunnett and Sobel compared their lower bounds to the actual P values for the important special case $\rho = \frac{1}{2}$ and where the h_i's have a common value h. For $\nu = \infty$ and (1.1) close to one, their comparisons suggest that the lower bounds are reasonably accurate for k=3, but for k=9 the accuracy diminishes considerably. They also examined the case $\nu = 5$. For k=3 the approximations were tolerable, but for k=9 the approximations were poor.

The approximation (3.5) consistently gave the most accurate results.

Table 1 shows the exact value of h so that P=.99, .95, .75, .50 for k=3,9. These values were taken from Dunnett and Sobel (1955). Included in the table are the values of h determined with (3.2) and (3.5) and (2.4). As can be seen, (2.4) nearly always improves upon both (3.2) and (3.5) without making any assumptions about the structure of the correlation matrix $\{\rho_{ij}\}$. For P close to one there is little improvement over the other approximations, primarily because (3.2) and (3.4) give fairly accurate results. As P decreases, though, (2.4) begins to give reasonably more accurate results.

Table 2 shows the approximation of P for ν=5, k=3,9 and various values of h. Again (2.4) nearly always improves upon (3.2) and (3.5), but unfortunately all three approximations are poor for k=9 unless P is close to one. Also observe that (2.4) is substantially more accurate for k=3 and P=.5.

3.2 Approximating a Distribution Occurring in Ranking and Selection

Let T_i (i=1,...,p+1) be p+1 independent random variables all having a Student's t distribution with ν degrees of freedom, and let $W_i=T_i-T_{k+1}$ (i=1,...,k). The joint distribution of the W_i's arises in the ranking and selection problem considered by Dudewicz and Dalal (1975). Table 3 shows the exact value of P (which was taken from the table in Dudewicz and Dalal) and the value of (2.4) for k=3,5; h=1,2,4 and ν=1,14,29. The value of (2.4) was determined using the table in Dudewicz and Dalal. As can be seen, the approximation does not always work well when ν=1, but otherwise it gives reasonably accurate results. Table 3 also includes an approximation based on the Bonferroni inequality $P\geq 1-\Sigma Pr(W_i>h_i)$ but as is evident (2.4) gives better results and in most cases the improvement is substantial.

3.3 Estimating the Probability of a Correct Selection in Ranking and Selection

For the final comparison, let $X_1,\ldots,X_{k+1}$ be k+1 independent standard normal random variables. Estimating the probability of a correct selection in ranking and selection problems requires evaluating

$$\Pr(X_1-X_{k+1}\leq h_1,\ldots,X_k-X_{k+1}\leq h_k) \qquad (3.6)$$

Evaluating (3.6) also plays a central role in Tong (1978).

Olkin, Sobel and Tong (1976) suggest a family of approximations of (3.6) that are based on majorization. To illustrate the accuracy of their approximation they consider k=5, h_1=3.2, h_2=2.7, h_3=2.5, h_4=1.9, h_5=1.7. The exact value of (3.6)is .8016. The closest approximation (in absolute value) based on their approach is .7802. If instead (2.4) is used, we get .8171. Obviously this one case is not a compelling reason to abandon the approximation proposed by Olkin, Sobel and Tong. It is difficult to make extensive comparisons because the quantity approximated by Olkin et al. is generally unknown. The point is that we have one more example where (2.4) gives good results.

Henery (1981) suggests another approximation of (3.6), which we compared to some of the exact values in Bechhofer (1954). For k=3 it worked reasonably well for $P\leq .8$, but unfortunately for $P>.8$ it gave very poor results and so it was not considered further (cf. Sathe & Lingras, 1980; Rice et al., 1979).

4. Summary and Concluding Remarks

In some instances the approximation (2.3) will give very accurate results, but as was illustrated this is not always the case. However it seems to usually give a reasonable approximation in most situations when k is not too large. Moreover, it is easy to use when the exact dis-

tribution is known for k=2, and so it may be useful in certain situations. More importantly, (2.3) appears to compare favorably to various approximation that have been proposed in the past, and it can give considerably better results when P is not too close to one. It is interesting to note that the Bonferroni inequality is known to usually give accurate results when P is close to one; (2.4) generally gives an even better approximation in these cases, but the improvement is not overly striking.

For distributions related to Student's t distribution, the comparisons made in Tables 1 and 2 suggest that (2.4) works tolerably well for k=5 and ν, the degrees of freedom, as small as 14. For k=3, (2.4) seems to even work reasonably well for ν=5. However, for k=9, all of the approximations considered here appear to be highly inaccurate except for a few cases where P is close to one.

Finally, no analytic results were given on the accuracy of (2.3), but the only analytic result concerning the other approximations is that they provide bounds for P. In some instances these bounds can be extremely inaccurate, in which case (2.3) might be considered. In fact, in terms of obtaining accurate approximations, the only motivation for preferring existing bounds is that they were invented first.

TABLE 1

Comparisons for the Multivariate Normal Case of Exact and Approximate Percentage Points, h, for Selected Values of P

	k=3				k=9			
P	(3.2)	(3.4)	(2.4)	Exact	(3.2)	(3.4)	(2.4)	Exact
.99	2.71	2.70	2.69	2.68	3.06	3.05	2.97	3.00
.95	2.13	2.09	2.05	2.06	2.54	2.51	2.29	2.42
.75	1.38	1.26	1.16	1.19	1.91	1.82	1.30	1.60
.50	.97	.70	.56	.59	1.59	1.38	.85	1.04

TABLE 2

Comparisons for the Multivariate t of Approximate and Exact P values for Selected Values of h, ν=5

	k=3			
h	(3.2)	(3.4)	(2.4)	Exact
4.21	.987	.989	.990	.99
2.69	.931	.944	.954	.95
1.32	.625	.721	.770	.75
.62	.139	.445	.515	.50
		k=9		
5.03	.987	.989	.998	.99
3.30	.903	.919	.997	.95
1.81	.415	.597	.944	.75
1.10	0	.269	.655	.50

TABLE 3

Approximations of Values Tabulated by

Dudewicz and Dalal

		k=3			k=5		
		Bonferroni (2.4)		Exact	Bonferroni (2.4)		Exact
ν	h						
1	1	0	.422	.402	0	.325	.285
14	1	.249	.552	.537	0	.483	.431
29	1	.265	.559	.545	0	.491	.440
1	2	.250	.572	.541	0	.519	.414
14	2	.724	.806	.798	.540	.776	.726
29	2	.745	.818	.811	.575	.788	.743
1	4	.557	.743	.711	.262	.750	.605
14	4	.983	.985	.985	.971	.980	.977
29	4	.989	.990	.990	.981	.986	.984

References

Bahadur, R. R. A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (Ed.), Studies in item analysis and prediction. Stanford: Stanford University Press, 1961.

Bechhofer, R. E. A single-sample multiple decision procedure for ranking means of normal populations with known variances. Annals of Mathematical Statistics, 1954, 25, 16-39.

Bechhofer, R. E., Dunnett, C. W., & Sobel, M. A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. Biometrika, 1954, 41, 170-176.

Dudewicz, E. J., & Dalal, S. R. Allocation of observations in ranking and selection with unequal variances. Sankhya, 1975, Series B, 37, 28-78.

Dunnett, C. W. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 1955, 50, 1096-1121.

Dunnett, C. W., & Sobel, M. A bivariate generalization of Student's t distribution, with tables for certain special cases. Biometrika, 1954, 41, 153-169.

Dunnett, C. W., & Sobel, M. Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's t-distribution. Biometrika, 1955, 42, 258-260.

Gupta, S. S. Probability integrals of multivariate normal and multivariate t. Annals of Mathematical Statistics, 1963, 34, 792-828.

Gupta, S., & Sobel, M. On selecting a subset which contains all populations better than a standard. Annals of Mathematical Statistics, 1958, 29, 235-244.

Henery, R. J. Permutation probabilities as models for horse races. Journal of the Royal Statistical Society, 1981, Series B, 43, 86-91.

John, S. On the evaluation of the probability integral of the multivariate t. Annals of Mathematical Statistics, 1961, 48, 409-417.

Johnson, M., & Kotz, S. Distributions in Statistics: Continuous Multivariate Distributions. New York: Wiley, 1972.

Krishnaiah, P. R. Multiple comparison tests in multi-response experiments. Sankhya, 1965, Series A, 27, 65-72.

McFadden, J. A. Urn models of correlation and a comparison with the multivariate normal integral. Annals of Mathematical Statistics, 1955, 26, 478-489.

Olkin, I., Sobel, M., & Tong, Y. L. Estimating the true probability of correct selection for location and scale parameter families. Department of Statistics, Stanford, Technical Report No. 110, 1976.

Rice, J., Reich, T., & Cloninger, C. R. An approximation to the multivariate normal integral: Its application to multifactorial qualitative traits. Biometrics, 1979, 35, 451-459.

Sathe, Y. W., & Lingras, S. R. A note on the inequalities for tail probability of the multivariate normal distribution. Communications in Statistics--Theory and Methods, 1980, A9, 711-715.

Tong, Y. L. Some probability inequalities of multivariate normal and multivariate t. Journal of the American Statistical Association, 1970, 65, 1243-1247.

Tong, Y. L. An adaptive solution to ranking and selection problems. Annals of Statistics, 1978, 6, 658-672.

Wilcox, R. R. An approximation of the k out of n reliability of a test, and a scoring procedure for determining which items an examinee knows. Psychometrika, 1983, 48, 211-222.

UNBIASED ESTIMATION IN A CLOSED SEQUENTIAL TESTING PROCEDURE

Rand R. Wilcox
Department of Psychology
University of Sourthern California

and

Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

Let p be the proportion of items within an item domain that an examinee would answer correctly if every item were attempted. This brief note provides unbiased estimates of p^t, for any integer t, when a closed sequential testing procedure is used.

Consider a single examinee, a domain of items, and let p be the examinee's domain score or true score. That is, p is the proportion of items in the domain of items that the examinee would get correct if every item were attempted. In some cases it is assumed that z, the number correct observed score, has a binomial probability function, and that for the population of examinees the distribution of p belongs to the beta family. This beta-binomial model has been used to solve many measurement problems (Lord, 1965; Lord & Novick, 1968; Wilcox, 1981a).

Let p_0 be a known constant, $0<p_0<1$. In criterion-referenced testing a common goal is to determine for every examinee whether $p \geq p_0$. Usually this is done by administering n items to every examinee and deciding $p \geq p_0$ if and only if $z/n \geq p_0$. Wilcox (1981b) pointed out that it is possible to improve uniformly on this procedure when computerized testing is feasible. The procedure is based on a closed sequential sampling scheme. This means that items are sampled and administered one at a time until an examinee gets m items right or M items wrong. In Wilcox (1981b) m was set equal to the smallest integer z such that $z/n \geq p_0$, and then M was set equal to n-m+1.

The purpose of this brief note is to provide unbiased estimates of p^t for any integer t, $1 \leq t \leq m$. It is noted that for t=1, an unbiased estimate is easily derived from results in Girshick et al. (1946).

After sampling terminates, let x be the number of items the examinee answers correctly, and let y be the number for which an incorrect response is given. The unbiased estimate of p^t is

$$\hat{p}^t = \begin{cases} \binom{m-t-1+y}{m-t-1} \Big/ \binom{m-1+y}{m-1}, & \text{if } x = m \\ \binom{M+x-t-1}{x-t} \Big/ \binom{M+x-1}{M-1}, & \text{if } y = M. \end{cases}$$

where $\binom{M+x-t-1}{x-t}=0$, if $x<t$.

To establish the above result, first it is noted that from Wilcox (1981b), the joint probability function of x and y is

$$f(x,y|p) = \begin{cases} \binom{m-1+y}{m-1} p^m(1-p)^y, & \text{if } x = m \\ \binom{M-1+x}{M-1} p^x(1-p)^M, & \text{if } y=M. \end{cases}$$

Proceeding as is done for the binomial case, it follows that $E(\hat{p}^t)=p^t$.

Henceforth, p^t will be written as $\hat{p}$ when t=1. The maximum likelihood estimate of p is $\hat{p}_m=x/(x+y)$. To gain some insight into how $\hat{p}$ and $\hat{p}_m$ compare, selected values of $E(\hat{p}-p)^2$ and $E(\hat{p}_m-p)^2$ were computed, and the results are reported in Table 1. As can be seen, $\hat{p}$ generally gives more accurate results than $\hat{p}_m$.

Two situations are briefly noted where unbiased estimates of p^t are important. The first is estimating the true score distribution. Suppose that for the population of examinees, p has a beta density given by

$$g(p) = \Gamma(r+s)/(\Gamma(r)\Gamma(s))p^{r-1}(1-p)^{s-1} \qquad (1)$$

where r,s>0 are unknown parameters. To estimate r and s, let $\hat{p}_i$ and $\hat{p}_i^2$ be the unbiased estimate of p_i and p_i^2, respectively, for the ith randomly sampled examinee, i=1,...,N. Proceeding as in Griffin and Krutchkoff (1971), it follows that

$$\hat{\mu}_t = N^{-1} \sum \hat{p}_i^t$$

can be used to estimate $E(p_t)$, where the expectation is taken with respect to the beta density. Thus, r and s can be estimated as described in Wilcox (1981a).

The second illustration has to do with the optimal linear estimator of p. Because $\hat{p}$ is unbiased, the linear estimate, $\tilde{p}$, that minimizes $E_p E(\tilde{p}-p)^2$ is given by $\tilde{p}=(\sigma_p^2/\sigma_x^2)(\hat{p}-\mu_1)+\mu_1$ where σ_x^2 is the variance of the marginal distribution of $\tilde{x}$, $\sigma_p^2 = \mu_2-\mu_1^2$ and $\mu_t=E(p^t)$ (Griffin & Krutchkoff, 1971). From the results given above σ_p^2 and μ_1 can be estimated yielding an estimate of $\tilde{p}$ (cf. Wilcox, 1978).

REFERENCES

Girshick, M. A., Mosteller, F., & Savage, L. J. Unbiased estimates for certain binomial sampling problems with applications. Annals of Mathematical Statistics, 1946, 17, 13-23.

Griffin, B. S., & Krutchkoff, R. G. Optimal linear estimators: An empirical Bayes version with application to the binomial distribution. Biometrika, 1971, 58, 195-201.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, R. M., & Novick, M. R. Statistical theories of mental test scores. Reading Mass.: Addison-Wesley, 1968.

Wilcox, R. R. Estimating true score in the compound binomial error model. Psychometrika, 1978, 43, 245-258. (a)

Wilcox, R. R. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 1981, 6, 3-32. (a)

Wilcox, R. R. A closed sequential procedure for comparing the binomial distribution to a standard. British Journal of Mathematical and Statistical Psychology, 1981, 34, 238-242. (b)

TABLE 1*

VALUE OF $E(\hat{p}-p)^2$ AND $E(\hat{p}_m-p)^2$

m	M	p:	.1	.2	.3	.4	.5
5	5		.0168	.0276	.0338	.0368	.0376
			.0126	.0236	.0350	.0448	.0489
5	10		.0083	.0143	.0198	.0256	.0307
			.0076	.0170	.0271	.0326	.0334
5	15		.0056	.0109	.0175	.0246	.0304
			.0059	.0154	.0236	.0282	.0308
5	20		.0044	.0098	.0171	.0245	.0304
			.0053	.0144	.0216	.0270	.0305
10	10		.0083	.0133	.0157	.0162	.0162
			.0071	.0120	.0155	.0190	.0208
10	15		.0055	.0088	.0106	.0122	.0141
			.0050	.0084	.0122	.0157	.0163

*The first entry in every cell is $E(\hat{p}-p)^2$, and the second entry is $E(\hat{p}_m-p)^2$.